DOCTOR OF HEALTH (DHEALTH)

Understanding users of a freely-available online health risk assessment

An exploration using segmentation

Hodgson, Corinne

Award date:
2015

Awarding institution:
University of Bath

Understanding Users of a Freely-Available Online Health Risk Assessment: An Exploration Using Segmentation

Corinne Susan Hodgson

A thesis submitted for a Professional Doctor of Health (DHealth)

University of Bath

Department of Health

December 2014

_____

Table of Contents

Tables and Figures

Acknowledgements

The author would like to thank two men who displayed great patience in the completion of this thesis. The first is my thesis supervisor, Dr. Gordon Taylor of the University of Bath. The second is my husband, Dr. Douglas Swallow.

The author would also like to thank the thesis examiners, Drs. Alan Buckingham of the University of Bath and Elizabeth Sillence of Northumbria University for their invaluable feedback. Finally, the author would like to thank the Heart and Stroke Foundation for access to, and permission to analyze, the Health Risk Assessment data, in particular Dr. Ahmad Zbib, Manager of Consumer e-Health.

Needless to say, the author would also like to acknowledge the over 120,000 consumers who gave consent for their information to be used for research purposes. Without their cooperation, this study would not have been possible.

Finally, I would like to dedicate this work to my father, Percy William Hodgson (1916-1982), who I think would have been pleased. *Quo fas et gloria ducunt.*

Abstract

Health organizations and governments are investing considerable resources into Internet-based health promotion. There is a large and growing body of research on health “etools” but to date most has been conducted using experimental paradigms; much less is known about those that are freely-available.

Analysis was conducted of the data base generated through the operation of the freely-available health risk assessment (HRA) of the Heart and Stroke Foundation of Ontario. During the study period of February 1 to December 20, 2011, 147,274 HRAs were completed, of which 120,510 (79.8%) included consent for the use of information for research and were completed by adults aged 18 to 90 years.

Comparison of Canadian users to national statistics confirmed that the HRA sample is not representative of the general population. The HRA sample is significantly and systematically biased by gender, education, employment, heath behaviours, and the prevalence of specific chronic diseases. Etool users may be a large but select segment of the population, those previously described as “Internet health information seekers.”

Are all Internet health information seekers the same? To explore this issue, segmentation procedures available in common commercial packages (k-means clustering, two-step clustering, and latent class analysis) were conducted using five combinations of variables. Ten statistically significant solutions were created. The most robust solution divided the sample into four groups differentiated by age (two younger and two older groups) and healthiness, as reflected by disease and modifiable risk factor burden and readiness to make lifestyle changes. These groups suggest that while all users of online health etools may be health information seekers, they vary in the extent to which they are health oriented or health conscientious (i.e., engaging in preventive health behaviours or ready for behaviour change). It is hoped that this research will provide other organizations with similar data bases with a model for analyzing their client populations, therefore increasing our knowledge about health etool users.

Abbreviations

BIC	Bayesian Information Criterion
BPAP	Blood Pressure Action Plan
BVR	Bivariate residual
CCHS	Canadian Community Health Survey
CI	Confidential interval
COPD	Chronic obstructive pulmonary disease
CVD	Cardiovascular disease
df	Degree of freedom
HEPA	Health-enhancing physical activity
HISB	Health information seeking behaviour
HSF	Heart and Stroke Foundation
HSFC	Heart and Stroke Foundation of Canada
HSFO	Heart and Stroke Foundation of Ontario
HRA	Health risk assessment
HWAP	Healthy Weight Action Plan
L^2	Model fit likelihood ratio chi-squared statistic
LCA	Latent class analysis
NPHS	National Population Health Survey
OR	Odds ratio
PAF	Population attributable fraction
RCT	Randomized controlled trial
RMPE	Recommended minimum effect
sd	Standard deviation
SE	Standard error
UK	United Kingdom
US	United States

1: State of Knowledge on Internet Health Information Seeking

As of March 2011, 30.2% of the world's population has been reported as using the Internet, with rates ranging from a low of 11.4% in Africa to a high of 78.3% in North America. Moreover, growth in Internet use is substantive: between 2000 and 2011 the percent of the population using the Internet increased 151.7% (i.e., by approximately 52%) in North America, to a high of 2,527.4% (i.e., 25-fold) in Africa (1).

In Canada, Statistics Canada reported the proportions of Canadians aged 18 and older using the Internet at least once a day has increased consistently over time, from 63.7% to in 2005, to 68.2% in 2007 and 75.1% in 2009 (2). By 2010, 80% of Canadian households had Internet access (3), increasing to 83% by 2012 (4). In the United States, the U.S. Census Bureau estimated that as of 2009, three-quarters of householders (76.6%) and households (76.3%) had access to the Internet from some location, with approximately two-thirds (68.7% of householders and 63.5% of households) having Internet access at home (5). In the UK, the Office for National Statistics reported that as of 2011, 82.9% of the population had ever used the Internet and 77% of households had Internet access, up from 61% in 2007 (6).

Computer-based communication has transformed the way people find and exchange information and its impact has extended throughout health care. As described by Gurak and Hudson, the use of Internet-based technology for health-related purposes (commonly referred to as ehealth) covers a broad range of consumer and health practitioner applications (7). In addition to clinical uses (e.g., telemedicine, electronic medical records, and electronic prescribing and consultations), ehealth can be used for educating both practitioners and patients. As noted by several commentators, the Internet has tremendous potential as a media or platform for health promotion and protection (8-11). It can not only provide information but, through interactive tools (etools) promote knowledge, attitudinal and behaviour change and provide a channel for social support (12). But who exactly uses the Internet for these purposes?

This chapter will look at our current state of knowledge concerning the use of the Internet for health purposes. It will begin by looking at health information seeking and then proceed to an examination of health etools. The objective is to summarize what we know about the type of people who use the Internet for health purposes.

Health information seeking

In many respect, health information seeking on the Internet can be viewed as a more recent form of health information seeking behaviour (HISB). A review of the literature on HISB by Lambert and Loisells in 2007 found no one dominant definition for HSIB (13); rather, its meaning has often been inferred from the behaviour itself and in some publications the term is used almost interchangeably with health consciousness (i.e., one seeks health information because one is health conscious). Lambert and Loisells argue that there are two main components in understanding the construct of HISB: the type of information (topic, level of detail and nature) as well as the method(s) by which it is retrieved (e.g., using search engines, discussion boards, or referrals) (13). At the same time, other aspects must be understood, including:

a) Personal or contextual antecedents that influence whether and how an individual responds to an information need, such as health consciousness or health concerns

b) Cognitive, behaviourial, physical and affective outcomes (e.g., increased knowledge, behavioural changes, change in health or wellness status, increased sense of control or empowerment) (13).

Health consciousness may be a predictor of health information seeking but what is meant by "health consciousness"? As noted by Hong (14), there has been little consistency between studies or over time in the definition or measurement of health consciousness. Introduced in 1988 by Gould in a marketing journal, the concept originally consisted of a set of health-related beliefs or attitudes. Based on survey data from 350 American university students, Gould argued that "health conscious" individuals are those who tend to be more evaluative of health care claims, more preventative in outlook, and somewhat more open to alternative or complementary therapies (15). In 1993, Kraft and Goodell (16) noted that Gould's Health Consciousness Scale (17) correlated positively with awareness of one's own health and some healthy behaviours but was not equivalent to the more recently-constructed and holistic theory of "wellness" that was becoming increasingly important in marketing and health care.

Writing in 1998, Jayanti and Burns argued that health consciousness should be "conceptually distinct from health motivation" but rather "refers to the degree to which health concerns are integrated into a person's daily activities" and the extent he/she is "wellness-oriented" (18). This suggests they saw health consciousness as reflecting behaviour (daily activities) and attitudes (wellness-orientation) but not necessarily as a driver or form of motivation.

In 2004, Dutta-Bergman defined being health conscious as neither a behaviour nor a single attitude or concern but rather as the concurrence of two specific attributes: a feeling of self-efficacy over health and an internal health locus of control (19). To Dutta-Bergman, health consciousness is an indicator of an intrinsic motivation for health, which may – but not necessarily – be actualized in behaviours such as actively seeking out information and resources or engaging in health-enhancing behaviours (19).

In 2009, Dissmore described being "health conscious" as a universal concern about health and well-being that can range in intensity (20). In this approach, a high level of health consciousness could be considered a positive attribute as it may be associated with greater intrinsic motivation for a healthy lifestyle and health-enhancing behaviours. However, although not discussed by Dissmore, there is also the potential that high levels of health consciousness may be counter-productive if combined with an exaggerated sense of susceptibility to disease or inaccurate perception of the seriousness of symptoms. As described by Wagner and Curran, such people may comprise the small minority (approximately 10%) of the "worried well" that frequently and, in the eyes of healthcare providers, inappropriately seek medical attention for a wide and fluctuating range of physical and psychological symptoms which cannot be traced to an underlying medical disorder (21).

It should also be recognized that in many cases the term "health conscious" has been used in the literature with little or no rigour. For example, a number of studies have defined or measured health consciousness in terms of a specific health-related behaviour such as food choices (22-25), preventive medical care (26) or drug use (27) -- even though such acts may be complex individual and social acts (28). In this approach, those who choose what are considered healthier behaviours are, *ipso facto*, labeled as being health conscious. This approach does not take into account the fact that, as indicated by results from population health surveys, different health behaviours may not be strongly related to one another (29). The person who might be labeled "health conscious" because of a behaviour such as jogging may still indulge in other activities that are in fact health detrimental, such as smoking or binge drinking. For example, principle component analysis of a variety of health-related attitudes and behaviours among a small (n=172) sample of older (ages 64-96 years) Americans found they formed into four groups based on health-related behaviours of information seeking, regular health routines, medical and self-examination, and risk avoidance. However, there was only modest association between these four groups of behaviours (30).

Newsom *et al.* argue that the lack of correlation between different health behaviours and attitudes shows there is no "single, health-consciousness motivation that

underlies all major health behaviours" (29). Rather, there may be what Bloch describes as a variety of "health behavior clusters," consisting of attitudes and beliefs concerning different domains such as health practices, safety practices, preventive health care, harmful substance avoidance, and environmental hazard avoidance (31). In other words, in addition to level of concern or awareness of health (20), individuals may vary in terms of which health behaviour "clusters" are important to them (31).

Another term that is often raised in discussion of health information seeking is health orientation. Roberts *et al.* describe health orientation as part of a more generalized set of personality traits ("health conscientiousness") characterized not only by health consciousness (awareness or concern) but also behavioural attributes such as good impulse control and the ability to plan, set goals and delay gratification (32). For Dutta-Bergman, health orientation is an intrinsic motivation for health-enhancing behaviours that is the result of being health conscious (19, 33). In one publication, for example, he defines health orientation as

> …a motivation-based construct, reflecting systematic individual-level differences within a population with respect to the extent to which individuals are motivated in health-related issues and are willing to process health-related information (34).

Elsewhere, he described health information orientation as having four indicators (39):

1) health consciousness, for which he uses Jayanti and Burns' definition based on the extent to which health concerns are integrated into daily activities (18);
2) health information orientation, which he refers to as "the extent to which the individual is willing to look for health information";
3) health-oriented beliefs; and
4) health activities.

Health-oriented activities are those that reflect not merely a high level of awareness or concern about health but also "a high level of active consumer responsibility and a willingness to engage in responsible actions" (33).

In 2009, DuBenske *et al.* (35) conducted confirmatory factor analysis of the Health Information Orientation Scale and found two unique and unrelated factors: 1) information engagement, which was formerly referred to as information seeking, and 2) information apprehension, formerly referred to as information avoidance. In a study including caregivers, DuBenske *et al*. found that those who scored high in information engagement

had greater information competence, defined as greater self-efficacy in their ability to understand and make use of health information. In contrast, those who scored higher in information apprehension had lower self-efficacy in their competency to seek, interpret and use health information (35). Other researchers have suggested that health information apprehension may be a function of not only information self-efficacy but a drive to avoid unpleasant or negative information or information that may demand changes in attitudes, beliefs or behaviours (36). The drive to protect oneself from potentially negative or demanding information may combine with an uncertainty orientation (drive to reduce uncertainty or tolerance of uncertainty) to shape one's coping style as a "blunter" (information avoider) or a "monitor" (information seeker) (36, 37).

At times, it appears as though the term "health orientation" is used in a broader sense. For example, in their study of 1,650 respondents to a mailed-survey, Wolff *et al.* (38) used a Health Information Seeking Orientation scale based on two attributes: 1) degree of autonomy in seeking health information and 2) amount of energy expended. Both attributes reflect behaviours (i.e., what people reported they did) rather than motivation. As a result, it is not clear to what extent this sort of health information seeking orientation reflects motivation for health-associated behaviours.

Another term that is relevant to HSIB is that of health conscientiousness. As briefly noted above, health orientation may be part of a personality trait ("conscientiousness") in the context of health matters ("health conscientiousness") (32). As described by Jackson *et al.* (2010), personality traits are assessed, and therefore at least partially, defined by behaviour, even though they are thought to refer to relatively stable patterns of not only behaviour but beliefs (thoughts) and attitudes (feelings) (39). Characteristics of conscientiousness include orderliness, industriousness, reliability, being more likely to follow social norms, more planful, more goal-directed and being better able to control impulse and delay gratification (39). Such personality traits support self-regulation, or the "process by which people seek to exert control over their thoughts, their feelings, their impulses and appetites, and their task performances" (p. 1773) (40). Self-control is critical in negotiating the temptations that surround us in everyday life, particularly when the goal is behaviour change or the adaptation of new behaviours (41).

There is considerable research showing that health and longevity are linked to a conscientious personality (32, 42). This link may be mediated by the positive relationships that have been observed between conscientiousness and educational attainment (43) and career success (44). There is also evidence that as people age, conscientiousness, or at least some aspects such as industriousness, impulse control and reliability (45), may increase and be positively correlated with increased attention to

preventative or health-enhancing behaviours (46). At the same time, it must be recognized that the relationship can be complex and mediated by the influences of social, environmental, and disease-related factors (42). For example, one study found the relationship between conscientiousness and adhering to medication adherence was not only mediated by willingness to follow a doctor's orders, which may be influenced by the quality of doctor/patient relationship, but stronger in older, as opposed to younger, individuals (47).

Health information seeking on the Internet

There is good evidence the Internet is being increasingly used for health information seeking. Increases in a broad range of online health information and advice seeking have been documented in:

- The US: According to surveys from the Pew Internet and American Life Project, Internet HISB increased from 33% in late 2001/early 2002 (48), to 40% in 2003 (49), 62% in 2007 (50), and 74% of Internet-users in 2011 (51). Proportions reported by The Harris Poll were even higher: from 71% of online adults in 1998 to 88% in 2010 (52). Moreover, 81% reported engaging in Internet HISB one or more times during the previous month, with a median of two and a mean of six times (52).

- Canada: Statistics Canada reported that in 2009 70% of those who were online in the home used the Internet to search for medical or health-related information, an increase from 59% in 2007 (53). This rate appears to be holding constant, with the 2012 Canadian Internet Use Surveying reporting 67% (54).

- The UK: As of 2011, 42% of UK adults who accessed the Internet in the last three months reported seeking health-related information (6), compared to 22% in 2008 (55).

At the same time, Schneider *et al*. argue there is evidence the Internet has yet to fulfill its potential for health due to sub-optimal dissemination (56). Although the proportion of the population using the Internet for health information is growing, a "digital divide" persists (57). Compared to the general population or those who do not use the Internet, studies have reported that those who use it for health-related purposes tend to be:

- Younger as opposed to older (48, 49, 58-72)

- Female (48, 49, 58, 59, 61, 65, 67, 70, 73, 74)
- More highly educated (48, 49, 58-60, 62-67, 69-71, 74-77)
- Higher income (57-59, 62, 65, 67-69, 76), although one population survey in the U.S. (49) and a small survey of cardiac outpatients in Canada (71) found no relationship. In the U.S., the relationship between Internet use and income may be confounded by insurance issues. In one American survey, for example, those with chronic conditions without health insurance had a rate of searching the Internet for health information 10 to 13 percentage points higher than those with insurance (48).
- Motivated by either their own medical issues or poor health (48, 49, 58, 61, 63, 64, 68, 70, 73, 74, 78) or those of others in their social network (61, 68, 70)
- Described as health-oriented or interested in health topics (79-81)
- In the U.S., white or Caucasian race (57, 60, 61, 65, 66, 75)

As noted by Dutta-Bergman, using the Internet to find health information is, like other communication behaviours, a goal-directed activity that is mediated by a variety of social and psychological factors (34). Moreover, as described by Moorman and Matullich, information acquisition and health maintenance may represent different types of motivation for seeking health information (82). Their analysis suggests that health information acquisition may not necessarily be sufficient to change behaviour: utilization of information (knowledge) will only occur if the consumer also has a health maintenance motivation or, in other words, has a health orientation (82).

As discussed, several studies have reported that poor personal health (48, 49, 58, 61, 63, 64, 68, 70, 73, 74, 78) or poor health of others in one's social network (61, 68, 70) may motivate health information seeking on the Internet. However, there are also studies reporting that Internet use for health purposes is higher among those with better, as opposed to worse, self-reported well-being (62, 66, 77). In some cases, the relationship between health and Internet use may be confounded by the quality of the practitioner/patient relationship: at least one study has reported more use of the Internet for health purposes among people who have less trusting and open relationships with their healthcare providers (64). In the U.S., the Internet may also be used to research health care providers (e.g., individual physicians, health management organizations, and hospitals) and insurance mechanisms such as Medicaid or Medicare (83).

Dutta-Bergman describes Internet health-information seekers as “more health conscious and health information oriented” (34). In a cross-sectional survey he found that people listing the Internet as a primary source for health information were “more likely to be health conscious, hold stronger health-oriented beliefs, and engage in healthy activities as compared to the respondent[s] that did not learn health information from the Internet” (84) (Pg. 284). Recent research found that among a sample (n=765) of Korean health mobile app users, structural equation modelling showed health consciousness had positive, albeit modest, relationships with not only health information orientation but also app use (85).

Other authors suggest that it may be premature to make conclusions about the motivators of Internet health-information seekers. They believe the development and testing of theories generalizable to all health information seekers have been limited by the tendency of studies to focus upon specific diseases (86) or dependent variables thought to be responsible for information seeking (82), or to see health-information seeking as a single and deliberate activity (87). As described by Boot and Meijman, Internet health-information seeking may involve more than the deliberate retrieval of facts: it may also be a means by which people try to reduce uncertainty, improve themselves (self-actualization), make social connections and/or entertain themselves (87). In fact, analysis of usage of one open-access or freely-available website (Daily Challenge) found that social ties were a significant predictor of etool engagement, such as return visits, opening emails sent by the site, and using an online self-reporting function (88). A 2011 survey using a nationally-representative sample of older Americans found positive associations between measures of social capital and Internet use (89). In other words, HSIB may be motivated by drivers other than simply the need for information.

A study of 1,016 adult women in New Jersey in the early 2000’s looked at the extent to which the use of the Internet for health information could be explained by health consciousness, health needs, or the costs of searching for health information (81). It found that although Internet HISB had a small but significant and consistent effect on behaviours thought to be indicative of health consciousness (such as diet, physical activity, smoking, and health screening), the relationship with health needs as indicated by diagnosed conditions was barely significant. Furthermore, factors reflecting the cost of seeking information, such as time and geographic barriers, were not significant. Nevertheless, in logistic regression, after accounting for the effects of age, education and income (which were significant independent predictors of Internet use), higher values on health and wellness behaviours, concern with health conditions and interruptions to work life were significant in determining Internet use (81).

More recent research suggests that Internet HISB can be a multi-faceted activity focused on addressing diseases, wellness, or a combination of the two (90). Weaver *et al.'s* 2006 survey of a relatively affluent online panel in the Seattle-Tacoma area found that only half of respondents reported seeking health information online, a proportion lower than the authors anticipated (90). Those who used the Internet to find health information did not constitute a "monolithic" population; rather, poor health status was associated with seeking out medical or disease-specific information whereas those in good health tended to seek out information on wellness and risk-reduction or prevention (90). In the case, of medical information, the findings of Weaver *et al.* conform to previous research suggesting that having a medical condition (83, 91), the number of conditions (92), level of anxiety about one's health (93), and having anxiety and high health self-efficacy (94) increases Internet HISB. However, not all studies have conformed to the pattern described by Weaver *et al.* In a study of a French-language Canadian site (*www.passeportsante.*net accessed 7/05/2013), even though the site was developed for illness prevention and health promotion, understanding a health problem or illness was reported to be the most important motivator for users (95).

Based on nine "valuegraphics" groups derived from a 15-item questionnaire on health care values and priorities, Wilkins and Navarro have argued that some of our assumptions about consumers may be incorrect (96). For example, their research suggests that while it is often assumed that most people care about improving their health, in reality close to 30% are not proactive and place a low value on maintaining or improving their health. Moreover, even among the two-thirds who are interested in health, their actions can vary significantly. Some groups (Independently Healthy, Ready Users and Naturalists) may strive for optimal health, while others (Family Centered and Loyalists) have lower expectations and at least one group, the Traditionalists, only act when a health problem presents itself (96). In other words, consensus has yet to be reached in the research literature on the characteristics and motivators of those engaged in Internet HISB.

Beyond health information seeking: interactive etools

As described above, substantive and growing proportions of the population are using the Internet to find and retrieve health information. As the Internet has evolved, it has developed the capacity to not only act as a platform for posting generic information (Web 1.0) but to interact with the user (Web 2.0). Health information seeking is therefore only one form of health-related behaviour possible on the Internet.

Are all health information seekers necessarily active users of Internet-based health tools (i.e., etools)? Seeking information is, in many respects, a limited and "flat" engagement with the Internet: the user enters terms in a search engine or search field and then decides whether to open, view, and/or print information. Etools, however, require a higher level of engagement from users. Typically, etools such as health risk assessments require inputting a number of data points; some, but not all, may also require completing a registration process.

Unfortunately, little has been published concerning the utilization of interactive health etools. A Pew Center survey conducted in 2012, for example, asked respondents to identify what they did online, such as seeking information to diagnosis a condition, investigate the safety of a medication or food, or research insurance issues, but there was no probing about the use of online health risk assessments or behaviour change etools (97). There is some evidence usage may be modest. A survey conducted in 2002 reported that only 24.7% of Internet health information seekers had used health behaviour, health promotion or disease management websites, with 62% stating they had no intention of using such programs (98).

Some etools are freely-available in that they are available to all Internet users without any related fee or cost. Binks *et al.* (40) refer to such sites as *ab libitum* programs. Other etools are sponsored by for-profit organizations and generate income through advertising (e.g., Sparkpeople *www.Sparkpeople.com*, accessed 7/05/2013) or membership fees (e.g., eDiets *www.ediets.com*, accessed 7/05/2013; Weight Watchers Online *www.weightwatchers.com*, accessed 7/05/2013, or Biggest Loser Club *www.biggestloserclub.com*, accessed 7/05/2013).

A short list of examples of freely-available interactive health etools includes:

- Health risk assessments: There are numerous health and/or disease risk assessments, as well as years of life calculators, on the web, sponsored by health organizations, governments, not-for-profits and for-profits. For example, open access cardiovascular disease (CVD) assessments include the American Heart Association's *My Life Check* (*http://mylifecheck.heart.org/*, accessed 7/05/2013), the National Institutes of Health National Heart, Lung and Blood Institute's ten-year coronary artery disease risk calculator (*http://hp2010.nhlbihin.net/atpiii/calculator.asp*, accessed 7/05/2013), European Society of Cardiology's *HeartScore* (*www.heartscore.org/Pages/welcome.aspx*, accessed 7/05/2013), International Taskforce for Prevention of Coronary Heart Disease PROCAM coronary risk assessment (*www.chd-taskforce.com/coronary_risk_assessment.html*, accessed 7/05/2013), World Heart

Federation's *Heart Age Tool* (*www.heartage.me/,*accessed 06/06/2014), Washington University's *Your Disease* Risk heart disease risk calculator (*www.yourdiseaserisk.wustl.edu*/, accessed 28/05/2014), National Health Service's *LifeCheck* and other self-assessments (*www.nhs.uk/Tools/Pages/Toolslibrary.aspx?Tag=Self+assessments*, accessed 7/05/2013), University of Nottingham's *QRISK*® calculator (*http://www.qrisk.org/*, accessed 7/05/2013), Patient.co.UK's *QRISK®2* cardiovascular risk assessment (*http://www.patient.co.uk/doctor/cardiovascular-risk-assessment,* accessed 7/05/2013), Project Big Life life expectancy and future hospital use calculators (*www.projectbiglife.ca*, accessed 06/06/2014) and the Mayo Clinic's heart disease risk calculator (*www.mayoclinic.com/health/heart-disease-risk/HB00047*, accessed 7/05/2013). It should be noted these sites represent only a brief scan of the freely accessible CVD assessments; there are also a wide range of cancer, diabetes, dementia and other online assessments.

- Portals or etools for entering, keeping and/or monitoring health information: Such etools can track blood pressure readings (e.g., *www.bplog.com*, accessed 7/05/2013), blood glucose readings (*www.glucosegraph.com*/, accessed 7/05/2013; *www.diabetease.com*, accessed 7/05/2013) or weight (*www.sparkpeople.com*, accessed 7/05/2013; *www.fitday.com*, accessed 7/05/2013; *www.weight-tracker.buddyslim.com*, accessed 7/05/2013; *www.myfitnesspal.com*/, accessed 7/05/2013). As well, many trackers have migrated to smart phones and are available at no or little cost as apps. A recent Canadian consumer study found that in 2012, 26% of cell phone users access health, wellness, fitness or nutritional information or tools through their devices (99).

- Microsoft HealthVault (*www.healthvault.com/ca/en*, accessed 7/05/2013) or HealthVault-compatible electronic health record portals: Sites such as *Heart360* offered by the American Heart Association (*www.heart360.org*/, accessed 7/05/2013), CardioSmart offered by the American College of Cardiology (*www.cardiosmart.org*/, accessed 7/05/2013), and the MyDoctor.ca portal operated by Practice Solutions in collaboration with the Canadian Medical Association (*www.mydoctor.ca*, accessed 7/05/2013) give consumers the capacity to store a wide range of personal medical and health information.

- Online tools for those who are caregivers of patients, such as the scheduling and networking etool *Lotsa Helping Hands* (*www.lotsahelpinghands.com*, accessed 7/05/2013) and caregiver stress self-assessments

(*www.agis.com/Document/5/caregiver-self-assessment.aspx*, accessed 7/05/2013).

Research on health etools

Over the past several decades, there has been a virtual explosion of publications and research concerning the Internet and health. For example, a search using the keywords ("Internet" OR "web") AND "health" using the Web of Knowledge database produced 72 citations for the period 1970-1979, 154 for 1980-1989, 4,775 in 1990-1999, and 34,995 for 2000-2009. A similar search using PubMed produced lists of 5,272,266 and 18,759 citations, respectively. Of course, these lists include any publication in which there could be reference to web-based technology. When the search was refined to ("Internet" or "web") AND "health promotion," the number of citations was much smaller but a similar trend over time was observed. Results for Web of Knowledge searches were none for 1970-1979, 1 for 1980-1989, 100 for 1990-1999, and 1,144 for 2000-2009. For PubMed, search results were no citations for the period 1970-1979, 1 for 1980-1989, 83 for 1990-1989, and 1,106 for 2000-2009. In other words, between the 1990s and the 2000s, there was an approximately ten-fold increase in the number of publications concerning health promotion and the Internet.

Research on Internet health etools has tended to focus on issues such as content (100-104), underlying theories of behaviour change that may be utilized (104-106), site architecture, functionality and design (107-111), methods of information searching or recruitment (112-114), and impact (115). As described by Danaher and Seeley (2009), research on etools may span the program evaluation continuum, from formative research (e.g., needs assessment) to process evaluation (e.g., operational efficiency), to the three stage of outcome evaluation: iterative intervention development and evaluation (Stage I), whether the intended effects are achieved in ideal conditions (Stage II efficacy research), or whether effects are achieved under broader and more realistic (i.e., real-world) conditions (Stage III effectiveness research) (116). Bennett and Glasgow (2009, p. 276) state that "few real-world (e.g., population-based) trials have been conducted" of web-based etools, with the majority of studies conducted in small and select samples (9).

The body of experimental evidence on health etools is considerable. Sufficient numbers of randomized controlled trials (RCTs) and quasi-experimental studies have been conducted to support meta-analyses of the efficacy of Internet-based interventions for smoking cessation (117, 118), sexual health promotion (119), patient empowerment (120), professional education (121, 122), health behaviour change (105, 123), alcohol consumption (124), weight loss and/or maintenance (125, 126), human immunodeficiency

virus (HIV) prevention (127), mental health treatment (128), and adult and childhood chronic disease management (129, 130) . Systematic reviews without meta-analysis have also been conducted of RCTs and/or quasi-experimental studies concerning web-based interventions for various health interventions (131), including those targeting depression and anxiety disorders (132-135), eating disorders (136), substance use (137), and nutrition, physical activity and/or weight management (138-141).

To explore the research paradigms utilized in studies of health etools, a review was conducted of publications concerning weight loss and/or maintenance over a ten year period, between 2001 and 2011. Weight loss was chosen as it represents one of the most common and popular forms of health behavior change attempted through the use of online websites, email, and instant messaging. Through PubMed and hand searches, 70 articles were identified that reported or concerned etools for weight loss or maintenance. As can be seen from Table 1, the majority of publications (52/70 or 74.3%) focused on determining the efficacy of weight loss/maintenance etools. Efficacy has been studied by using empirical methods such as RCTs (34/52 or 65.4%), quasi-experimental studies such as pre- and post-testing (9/52 or 17.3%), meta-analysis (3/52 or 5.8%), or by reviewing the quantitative literature (7/52 or 13.5%). These findings are not atypical: a 2013 review of cancer prevention and control etools by Sanchez *et al.*, of which many addressed common modifiable risk factors such as diet, weight and physical activity, reported 86% were designed to test efficacy, with 88% using RCTs methodology (142).

Table 1: Analysis of Weight Loss/Maintenance eTool Publications, 2001 – 2011

Research Focus	Type	Number (%)	Citations
Etool efficacy	Systematic review with meta-analysis	3 (4.3%)	(125, 126, 143)
	Review without meta-analysis	7 (10.0%)	(140, 141, 144-148)
	Randomized controlled trials	33 (47.1%)	(121, 149-180)
	Quasi-experimental/uncontrolled trials	9 (12.9%)	(101, 181-187)
Other etool aspects	Etool features or content	6 (8.6%)	(138, 146, 188-190)
	User characteristics	10 (14.3%)	(153, 191-200)
	Other (editorial or expert opinion)	2 (2.9%)	(201, 202)
Total		70 (100.0%)	

As noted by Bennett and Glasgow (2009), most experimental and quasi-experimental studies have recruited samples of convenience either on- or off- line (9). Specific populations are targeted, such as soliciting:

- the general public through newspaper advertisements or other mass media (151, 156, 157, 160, 168, 171, 173, 176, 177, 187)

- patients of primary care practices, health maintenance organizations (149, 161, 178, 200) or hospitals (170)
- health insurance enrollees (153, 166)
- university faculty, staff and/or students (162, 167)
- employees through workplaces (121, 159, 172, 193, 197, 198, 203)
- members of pre-existing online research panels (180) or a research participant database (150)
- church members (201)
- members of a previous or ongoing weight-loss program, sometimes through newspaper advertisements (152, 182, 183)

Unlike open-access websites without inclusion and exclusion criteria, "efficacy trials typically limit reach by seeking motivated, homogeneous participants with minimal or no complications or comorbidities" (p. 1263) (204). In the review of weight management etools, some studies selected for those with a specific condition, such as hypertension (149, 200), diabetes (153), dyslipidemia (200), or heart disease (153), whereas others may exclude those with chronic conditions (169).

Perhaps the most common difference between open-access, totally online etools and those tested in RCTs has been geography, in that many studies require attendance at in-person clinics, appointments or training sessions at baseline and/or follow up (149, 151, 152, 156, 157, 159, 161, 162, 164, 169-173, 175-177, 183, 187). In fact, a recent review of 83 online interventions found that 76% required participants to interact with counselors (205).

In efficacy trials, as with other types of research using the RCT design, researchers may use a variety of strategies to motivate and retain participants, ranging from incentives to individual case management (206). In this review of weight loss etools, for example, six offered incentives to participants (149, 171, 193, 195, 207).

Estimating the enrollment rate in RCTs of weight loss etools is often difficult, as recruitment may be conducted among the general public using mass media advertising (150-152, 156-158, 173, 175, 176, 182, 186, 187) or among unspecified numbers of patients in health care settings (161, 166, 170, 178, 200), employees at workplaces (159, 169, 172) or students and/or employees at universities (162). When enrollment rates are reported, they are often low. For example, Glasgow *et al.* report that of 79,378 Kaiser-

Permanente patients, 1,402 or 1.8% participated in an RCT (153). For a physical activity program, Buis *et al.* report that about 16% of eligible university subjects volunteered to participate, of which about 21% failed to log on even once (208). Likewise, Couper *et al*. report that of 28,460 adults invited to an online nutrition tool, 15% visited the site and 8.9% enrolled (209). Uptake of the same etool can vary between settings: in one study, completion rates for a brief online health survey ranged from 30% to 95% between work sites, with enrollment for a weight loss program ranging from 17% to 49% (193).

Given what is known about enrollment rates, it is not surprising that RCT samples are generally not representative of the general population but tend to reflect what is known about Internet health information seekers. For example, in recruiting overweight and inactive adults for a study of an online etool, Anderson-Bill *et al*. found that those who responded to online and print advertisements and consented to enroll in the online study tended to be middle-aged, well-educated, upper-middle class women whose unhealthy behaviours put them at increased risk of obesity and obesity-related chronic conditions (210). Likewise, an Australian study reported that 62% of those completing an initial assessment for an online QuitCoach smoking cessation program were female with a median age of 34 years (211). However, targeted marketing and recruitment methods can influence the demographics of respondents. In three studies, for example, online recruitment was able to attract larger-than-expected proportions of non-White (113, 212, 213), male (212) and/or less-educated (213) participants.

Even within the controlled environment of RCTs, adherence with online resources can be low (145, 160). Attrition rates typically range between 20% and 30% (151, 152, 165, 169, 182, 187), although one review suggested it may be as high as 50% (205). In addition to incentives, prompts and reminders are frequently used to promote adherence (200).

There is evidence that enrollment and retention rates may be even lower for non-experimental or freely-available sites. For example, Wanner *et al.* compared engagement with a physical activity online tool between open access participants to those recruited for an RCT (214). Among the open access participants, 4.8% of first visits results in registration and of those, only a quarter (25.8%) visited the site repeatedly; in contrast, 67.3% of RCT participants visited the site repeatedly. In another study, recruitment among members of a U.S. health system was 4.3%, ranging from 1% to 11% depending upon the type of incentive offered (112). Even among those already “on the web,” etool uptake may be low. A study in Denmark of physically inactive adults who had completed a health survey online (n=12,287) reported that of 6,055 randomly chosen and given access to an activity-promoting website, an appreciable proportion (42%) was lost to

follow up. Of 3,156 participants who were followed, less than a quarter (22.0%) logged onto the website even once and even fewer (7.0%) logged on frequently (215). In a study of a health risk assessment offered to Dutch employees, the participation rate was 33.7% (2,289/6,790), of whom only 637 (27.8%) completed a program evaluation survey (216).

The literature also suggests attrition rates for non-experimental weight loss etools are high: 80% for a freely-available and anonymous 15-week program (182), 93.9%, 87.8% and 83.3% for, respectively, three-, six- and 12-month memberships for a low-cost site sponsored by a Swedish newspaper (185), and up to 90% for both a 12- and 52-week commercial program in Australia (181). Moreover, as found in RCTs, users varied in the extent to which they interacted. Johnson and Wardle's retrospective study of 3,621 users of a commercial online program who entered at least two weights over a period of at least 28 days (i.e., were at least minimally engaged) found some members visited most days but others only occasionally (184). A small study of the first 204 overweight-to-obese adult users of a freely-available weight loss program reported that about half completed the self-assessments and less than a quarter utilized tools provided, such as an activity log, journal, weight tracker, or meal planner (191).

Health risk assessments

As discussed, there are several freely-available health risk assessments on the Internet but to date the research literature has focused largely on the validity or reliability of the risk calculations (217, 218). Although little data have been released concerning the uptake of freely-available online health risk calculators, there is some evidence suggesting a relatively low participation rate, with reports of 22.4% for online risk assessments offered to members of an American group health plan (219), 10.6% for a disease self-management etool advertised to primary care patients (220) and 5.2% among university employees for a diabetes risk calculator (221). In such studies, users tended to be predominantly female and middle-aged (219-221).

Four studies have been published that provide more detailed information on the type of people who use online health risk assessment. The first is a study of an online German diabetes risk assessment which over a six-month period (March to August, 2007) attracted 32,055 unique visitors (222). Of the unique visitors, 28,564 (89.1%) started the assessment and 24,844 (77.5%) visited the "score" page at its end (i.e., completed the assessment). Of 24,453 complete records available for analysis, the mean age was 48 years (sd=10), with 44% of users being female and 56% male. As discussed by the authors, the distribution by gender was unusual for health etools, suggesting the audience "may be different from regular online health information seekers" (pg. 111)

(222). Interaction with the site was relatively brief: 52.8% of visitors remained on the site between two and 15 minutes and 30% left within the first two minutes (222).

The second article is Brouwer *et al*.'s 2010 study of users of a freely-available Dutch heart health assessment (223). Over a three-year period, there were 285,146 unique IP visits to the site. Although half of visitors left the site within 30 seconds, 81,577 (28.6%) completed the registration process. Compared to the general Dutch population, registrants were more likely to be female, younger, and more highly educated; they also had a lower smoking rate than the general population (18.7% vs. 29.6%), fewer were overweight or obese (30.5% vs. 46.5%), and more complied with saturated fat intake recommendations (63.2% vs. 10.0%). At the same time, registrants were less likely to meet the recommendation to be physically active five to seven days a week (42.2% vs. 55.0%). In regression analysis, women, visitors aged 40 to 50 years, those with a medium education level and with a normal BMI (between 18.5 and 25.0 kg/m^2) were more likely ($p<.05$) to start and finish the physical activity and saturated fat intake modules (223).

The third study looked at engagement with online health risk assessments among members of a group health plan based in Seattle, Washington (219). Of the approximately quarter (22.4%) of eligible patients who accessed the etools, the majority were female, middle-aged (41-65 years), to have had a recent well-care visits, and were less likely to be smokers or to have depression or hypertension (219). This characteristics suggest uptake was by health conscious and perhaps even health conscientious patients.

The fourth study relevant to this thesis was released after the completion of the *viva voce.* This publication presented analysis of a large amount of data (approximately 2.7 million records from 13 countries) collected through the freely-available Heart Age calculator (224). Based on self-report of age, gender, parental history of heart problems, pre-existing heart attack, stroke, rheumatoid arthritis, chronic renal disease, atrial fibrillation or diabetes, ever smoking, height, weight, total and high density lipoprotein, and systolic blood pressure the authors concluded the tool reached users with low-to-moderate CVD risk (224). A comparison between the Heart Age and the HRA populations is difficult, however, due to differences in the type of data collected. The HRA, for example, does not ask for cholesterol or systolic blood pressure measures while Heart Age does not include questions on dietary behaviours, level of physical activity, salt and alcohol consumption, perceived stress, marital status, education, occupation, or readiness to change behavioural risk factors.

In addition, although not a health risk assessment, Lemire *et al.* (95) have reported on users of a freely-available Quebec-based health information website,

*www.Passeportsante.*com (accessed 7/05/2013). Of 2,923 users, two-thirds were women, 93% were aged ≥30 years, with 11% >65 years, and 57% reported consulting other, similar health-concerned web sites (i.e., were engaging in Internet HISB) (95). Factors that explained 35% of variance in frequency of use of the site were, in order of importance, perceived usefulness of the site, importance of health information found in the print media, level of concern about health, importance attached to opinions of physicians and other health professionals, trust in the information, and gender (95). Age, user-friendliness of the site and quality of information did not influence frequency of use.

As noted, there are numerous health risk assessments on the Internet. But what do we know about the effectiveness of risk assessments? A 1987 review by Schoenbach *et al.* of paper-based assessments found few indications they changed health beliefs or behaviours and only limited evidence quantitative risk messages such as risk scores had any effect on users (225). More recently, in 2010, the U.S. Task Force on Community Preventive Services reviewed the evidence concerning worksite-based health risk assessments (226). Studies included 32 stand-alone health risk assessments (of which only eight delivered computerized feedback) and 51 programs in which the assessment was combined with additional interventions. Stand-alone risk assessments were found to have small and inconsistent effects on behaviour, whereas for combined interventions there was sufficient evidence of impact for six behaviours (smoking, alcohol and seatbelt use, dietary fat intake, blood pressure and blood cholesterol) and three outcome indicators (health risk estimates, absenteeism and healthcare utilization) (226). The authors concluded there was sufficient evidence that a health risk assessment with feedback "has utility as a gateway intervention to a broader worksite health promotion program that includes health education lasting at least 1 hour or being repeated multiple times during 1 year and that may include an array of health promotion activities" (p. S257-8) (226). In other words, this research suggests that health risk assessments need to be part of broader health promotion efforts if they are to do more than merely educate.

The efficacy of the Heart Age calculator compared to the Framingham-based REGICOR risk score in improving modifiable CVD risk factors was the subject of a Spanish RCT in 2014 (227). However, the study is not a good representation of the operation of a freely-available health etool. First, all participants had to volunteer for the study, give informed consent and were aware they had become members of a group (control or experimental). These factors shift the act of completing an online risk assessment from being private and for self-assessment to the domain of a public action for the purpose of fulfilling a commitment as a member of a group. Second, all participants had to attend in-person baseline and follow-up appointments where anthropometric (height, weight, abdominal waist circumference) and biologic (blood

pressure and a blood sample drawn for analysis of lipid profile) measurements were taken. Third, as part of the research protocol, those completing either the Heart Age or REGICOR assessment had "their risk value … communicated and explained to them" by the researchers (227). As a result, even though both the Heart Age and REGICOR intervention groups demonstrated significant decreases in their risk scores at 12-month follow-up compared to the control group (227), the findings may not be generalizable to the open-access setting. The effectiveness of Heart Age, or other risk calculators, outside of experimental settings remains unclear.

A recent (2014) qualitative study in Australia used a "think aloud" methodology to assess attitudes and beliefs of 26 primary care patients as they completed two online CVD risk assessments (228). The primary objective of the study was to compare reactions to different ways of presenting CVD risk information ("heart age" compared to absolute 10-year risk) but the authors noted "an interesting paradox: online heart age calculators are easily misunderstood and the results may be dismissed if the information is unexpected or negative, but the process of using such calculators may motivate lifestyle change regardless of the outcome" (228). This effect may be due to the ability of online health risk calculators to prompt people to consider the effects of their behaviour on their health and thus the need for change (228). Whether these considerations lead to change may depend upon the individual's intentions and readiness to change. In their study of mobile app users, Cho *et al.* distinguished between information- and behaviour-oriented users (85). The former used apps such as *WebMD* to search for information (e.g., symptoms or medications) while the latter used monitoring and health management apps to change or maintain behaviour (e.g., physical activity, diet, blood pressure or blood glucose) (85). It may be that health risk assessment users vary in a similar way, in that only some have the motivation necessary to move beyond information-gathering to behaviour change.

Summary

Over the past several decades, there has been dramatic growth in many countries in the number of adults who are using the Internet as an alternative means of engaging in health information seeking. Population surveys suggest that such people tend to be female, younger rather than older, more highly educated, and to be health conscious (i.e., to be concerned about health and well-being) or to have health concerns. Thus, people who seek health information online may be presumed to be not only health conscious but health-oriented, in that they are actively engaged in HISB. What is less clear is the extent to which Internet health information seekers are engaged in behaviours beyond just HISB,

i.e., are willing or able to commit to health-promoting behaviours or to avoid health risks. Conscientiousness, a personality trait characterized by compliance with social norms, self-regulatory skills (the ability to delay gratification and control impulses), industriousness and orderliness has been found to be positively linked with health, possibly through its positive relationships with education and career success.

The capacity of Web 2.0 means the Internet empowers consumers to do more than merely seek out "flat" information but to interact with electronic health tools for health risk assessment, disease self-management, and/or behaviour change. In response, there has been a dramatic growth in the amount of experimental research on health etools. However, to date research has focused primarily on design and efficacy using experimental methods such as RCTs. As a result, there was been relatively little focus on external validity and the generalizability of etools (142).

In addition, although there has been research on Internet health information seekers, little has been published about the use of risk assessment or health behaviour change etools. Those creating health etools may assume that people who engage in Internet HISB are also interested in health-enhancing behaviours or behaviour change. But those who are information-oriented need not be health behaviour-oriented or ready to engage in health-enhancing behaviour.

Because of the paucity of research on freely-available etools, investments into the development of *ab lbitum* etools cannot be based upon empirical information. Rather, resources tend to be developed based upon assumptions about what types of people may comprise the "black box" of potential users. Buist *et al* argue that in the case of health plans, understanding users of electronic health risk assessments (eHRAs)

> … is relevant for several reasons. First, if eHRAs are to be used to characterize the health status of enrolled populations, it is important to understand how individuals who complete these assessments differ from those who do not; without this knowledge, health systems could make a biased assessment of the health status of their covered populations and could poorly target resources. Second, understanding selection factors for completion will be critical for accessing whether use of these tools leads to improved health outcomes and population. Finally, characterizing individuals who do not complete these tools provides an opportunity for reaching broader audiences for higher completion rates (219)

Although Buist *et al.* are working in a health plan environment, similar challenges are faced by health charities, government agencies or other organizations creating and operating open-access online health risk assessments. More and better information about the users of freely-available etools, preferably through ecologic research conducted with "a real-health promotion program rather than a laboratory-based experiment" (229) is needed to open up the "black box" and learn more about who uses them and their needs,

as well what segments of the population may be missed. If some of these “missed” segments are of high priority, special strategies may be required to optimize uptake of the etool, such as promotions or targeted marketing campaigns (e.g., using ethnic media outlets) or collaboration with primary care providers (230).

Real-world, observational research typically gives the researcher little or no control over exposure, measures and subjects, thus limiting the ability to determine causation (231). Nevertheless, this type of research would be a valuable first step in understanding how health etools operate in uncontrolled settings.

2: Research Objectives

Considerable numbers of not-for-profit, government and other organizations have invested, and continue to invest, substantive resources into the development and operation of open-access health etools, such as health risk assessment. Despite the fact that etools often have the capacity to capture data on usage and users, as discussed in the previous chapter, relatively little information has been published on open-access health etools (181, 182, 185, 191, 214) or health risk assessments (219, 222, 223, 232). A number of factors may be involved. For example, those creating freely-available etools may be health promoters with limited resources for, or interest in, analyzing data captured during the course of operating a program. There may also be a reluctance by some organizations to put information that may be considered proprietary into the public domain or ethical concerns about disclosing personal health information.

The goal of this study is to provide insights into the type of people who utilize health etools in the "real world," as opposed to samples created for experimental settings by analyzing the data base created by the Heart and Stroke Foundation's (HSF's) Health Risk Assessment (HRA). As noted by Weaver *et al.*, relatively little research has explored differences within samples of Internet health information seekers (90); as a result, such populations are often treated as though they are homogeneous or monolithic (233). Thus, one of the objectives of this study is to determine the validity of this perception

It should be noted that this study is a form of data mining or knowledge discovery in databases, i.e., the analysis of data sets, particularly large databases, in order to extract new findings or to summarize the data in potential new and useful ways (234-236). Data mining is perhaps best known for its application in commerce, such as tracking customers' purchasing patterns in order to guide purchasing or marketing decisions (236).

The HRA database does not fall into the domain commonly referred to as "big data," in that the size of the data base is within the capacity of common or typical software packages and does not require specialized tools (237). In addition, the range of data available for this analysis is more limited than in most big data scenarios, where there is not only a huge volume of data but data of various types (e.g., structured, semi-structured and unstructured) that may arrive at different rates or velocity (238). Rather, in this study an emphasis will be placed on statistical procedures that most organizations with similar

databases generated by online health etools can readily access, such as those included in standard commercial packages such as SPSS.[1]

The focus on segmentation procedures stems from two observations by the author. First, there is great interest among policy-makers and health promoters in psychodemographic segmentations, as they are seen as providing helpful insights into consumer and/or political behaviours and attitudes (239). However in the author's personal experience working with public opinion polling firms conducting segmentation analysis for clients, it is typical for only one segmentation solution to be described. This also occurs in some of the published literature describing segmentation (240, 241), although other authors describe the process by which they selected one solution over alternatives (229, 242, 243). What policy-makers and health promoters may not appreciate is that not only can the same segmentation procedure produce alternative numbers of groups, but alternative procedures can produce different groupings (244-246).

Research approach

In traditional health sciences research, the focus is upon developing general laws that explain phenomenon, such as the effect of a medication or risk factor on the course or development of a disease. In this approach, the focus is often upon the testing of hypotheses and determining the statistical significance of specific relationships between variables. However, the research for this thesis will utilize a critical realist approach in that it will be based upon the concept that statistical procedures are not accurate diagnostic tests but rather tools that can be used in various ways to further our understanding of an independent and ever-changing reality (247). Empirical methods are utilized but it is important to recognize that they are constrained both by the number and type of observations captured and by the analytic methods chosen by the scientist. Thus, the knowledge generated by science is not fixed but, as it is dependent upon a number of factors, can be described as a transcendental reality (248).

The following table summarizes the domains or levels of reality as described in critical realism and its correspondence in the proposed analysis of the HRA database. It reflects the fact that HRA users represent a variety of people who choose to come to the site for various personal, medical and/or societal reasons. Although the HRA is in itself structured, with set questions and response options, users have the liberty to interact with the system in any way they choose. They may, for example, start to complete the

[1] SPSS (Statistical Package for the Social Sciences) has been renamed by IBM as PASW (Predictive Analytics Software Package) . However, as it is still commonly referred to in past and present publications and resources as SPSS, this name will be used.

structured set of questions but have control over whether to respond honestly and may quit at any point. Thus, HRA use is in many respects a complex activity, the mechanisms of which can only be vaguely discerned by the records (the actuals) left behind by some – but not all – users.

In this analysis (Table 2), true reality is the intransitive universe of all interactions with the HRA, including visits which do not result in the completion of the questionnaire. In contrast, the HRA data base for analysis represent the level of the actuals, in that it is a record of responses and activities made by a sub-sample of all users, i.e., those who completed the assessment. This database includes both users' responses to HRA questions and web metrics routinely captured by the system, such as various time stamps. In many respects, this database is a vast and unorganized "data dump."

Table 2: Critical Realism Domains and Corresponding Levels in the Proposed Research

Domain	**Definition in Critical Realism** (247, 248)	**Level in HRA Research**
Reality	True reality (intransitive), consisting of events, experiences and the underlying mechanisms	Universe of all HRA user visits, including those who do and do not start or complete the HRA.
Actuals	Events and experiences (i.e., what is happening), whether or not we observe it	Raw database of all HRA data points and web metrics for those users who complete the HRA.
Empirical Events	The transitive or observable data	Research database created from raw data, which is analyzed by statistical procedure selected by the researcher

To be interpreted, this raw data must be captured and organized into an analyzable data file (i.e., the research database). This requires making choices about the variables to be captured, as well as the type of records to include. As described by Fayyad *et al*., in data-mining statistical analysis or "the application of specific algorithms for extracting patterns from data" refers to only part of the process that should be utilized in the analysis of large, non-experimental databases (249). To achieve what is referred to as knowledge discovery in databases or meaningful analyses, a number of additional steps are required, such as deliberate and careful data preparation, selection and cleaning and informed analysis and interpretation rooted in prior knowledge or theory

(249). If these steps are ignored and statistical testing is conducted blindly on a raw database, results could be meaningless or even misleading; hence, “data dredging.”

To interpret the empirical data in the (organized) database, a variety of statistical procedures will be utilized. Statistical procedures are, in themselves, “dumb” tools and the choice of procedures and how results are interpreted reflect the choices of the researcher. This fact will be most evident in segmentation of the research database: as will be described, there are various options for segmentation that may produce different results. Such groupings are not tangible or “real” groups but, as will be discussed, representations of reality. Segmentations can vary in terms of how robust they are and their face validity, but they may also differ in how useful they are in giving organizations new information and insights.

In keeping with a critical realism approach, the research will consist of several phases.

- **Description of observables**: In this case, the observables are health assessment responses and website usage data. Responses within the health assessment (i.e., user responses) will be analyzed to show general demographic and health characteristics of the population (Chapter 5).
- **Analytic resolution:** The focus of this stage is to “separate or dissolve the composite and the complex by distinguishing the various components, aspects or dimension” (247). In the case of the HRA, the goal will be to move beyond the gloss of all HRA respondents by showing how the sample varies from other populations (Chapter 6) and looking for segments or sub-groups (Chapter 7).
- **Abduction/theoretical redescription and retroduction:** This phase consists of developing and testing different models (i.e., segmentation solutions) in order to better understand the essence of essential properties of the population (Chapter 7).
- **Comparison of models:** Different segmentations will be compared to determine which may be more useful in explaining etool users’ behaviour (Chapters 7 and 8).
- **Concretization and contextualizaton:** In critical realism, this phase is typically devoted to interpreting how different structures and mechanisms interact at different levels, under specific conditions, or as concrete events and processes. In this analysis, this phase will consist of discussing how findings from the HRA analysis could be used by other organizations with similar data bases (Chapter 9).

Research questions

The objectives of this study are to:

1) Describe the HRA population so as to better understand the type of people who utilize an open-access, freely-available online health risk assessment, thereby filling a gap in the current literature (Chapter 5)
2) Compare the HRA population to other samples (Chapter 6) in order to determine:
    - How the self-selected sample of Canadian HRA users varies from the general population of Canada
    - Whether the use of an incentive has a significant impact upon the type of users who complete the HRA
    - To what extent users of the open-access HRA are similar to, or different from, samples recruited for etool RCTs
3) Challenge the assumption that open-access etool user population are monolithic by conducting exploratory segmentation using available HRA data points (Chapter 7)
4) Show further etool engagement by the HRA population and determine whether segments are helpful in understanding who does or does not enroll or interact (Chapter 8)

As well, it is hoped that this research, particularly the work on segmentation, could act as a model for other organizations and stimulate greater publication and sharing of information on open-access etool utilization.

Ethics

Consent for the research has been given by the Heart and Stroke Foundation of Ontario, as documented by a written Memorandum of Agreement. The proposal has also been approved by the University of Bath School for Health Research Ethics Approval (SREAP).

Steps taken to ensure the research was conducted in an ethical manner include:

- Submitting the project to an ethics review board and obtaining approval;
- Excluding from the research database all records which were not completed by the individual for him/herself (i.e., was completed for someone else or to review the site) and for which the user did not indicate consent for the use of de-identified information for research purposes

- Excluding from the data available for download by the researcher all fields that might identify participants, such as email address, IP address, username and password.

The privacy of participants was protected by ensuring that all records were identified only by a system-generated identification number. Data were analyzed in aggregate so information specific to any individual or record would not be divulged.

Summary

A number of organizations are now operating freely-available online health risk assessments capable of capturing data on large numbers of people but to date little information about these populations has been published and put into the public domain. The objective of this research is to fill this gap by conducting an analysis of one such data base. In the process, four activities are to be undertaken: 1) description of the HRA population, 2) comparison to other populations, 3) exploratory segmentation, and 4) analysis of follow-up data to test the utility of the chosen segmentation. It is hoped this research may provide models for analysis and hypothesis for further testing by organizations operating freely-available health etools.

3. The Health Risk Assessment Research Database

In describing the creation of the HRA research database, it is helpful to understand its context. Thus, a description is given of the HRA and its development (history) over time. This will be followed by a description of how the research database was formed.

History and description of the HRA

In 1999, the HSF collaborated with the Ontario Ministry of Health and Long-term Care (MOHLTC) in developing a five-year stroke strategy. This strategy involved re-organization of stroke pre-hospital care, acute care, rehabilitation and secondary prevention by the MOHLTC, as well as the development of public awareness-building and primary prevention programs by the HSF. As part of the stroke strategy, the HSF started to develop programs and initiatives to address the issue of hypertension, one of the most important risk factors for stroke. Literature at the time suggested that rates of undetected and/or untreated hypertension were unacceptably high in the general Canadian population (250).

In 2001, as part of the population-based hypertension strategy, the HSF supported the development of online health risk assessments. Initially, three online assessments were created: one to assess cardiovascular risk, another to assess hypertension risk, and the third to assess patient's quality of hypertension management. In developing these assessments, the goal was to avoid a "checklist" approach which would provide most participants with little or no new or helpful information (e.g., someone who is obese is probably already aware of the fact and cognizant of the health risks associated with obesity). Rather, the assessments were designed incorporating the Transtheoretical Model of Change (251). In the case of a participant who is obese, for example, a follow-up question would establish his/her readiness to make changes. Based on existing models, the follow-up question was phrased in the following manner:

When would you be willing to make changes to [insert behaviour or risk factor]*?*

( ) In the next month [in 2011 changed to read "in the next 30 days"]
( ) Within the next 6 months
( ) I'm not planning to make changes

The follow-up question made it possible to tailor risk factor messaging to the participant's readiness to change. Thus, someone in the preparation stage (i.e., ready to change in the next month/30 days) would be given a message with information on how to start making changes, someone in the contemplation stage (i.e., thinking of making

changes within the next six months) would be given advice on how to prepare for future change, while someone in the precontemplation stage (i.e., not willing to make changes) would receive a supportive message to "keep the door open" to future consideration of change. The objective was to meet participants at their stage of change so as to minimize the chances of alienating them or providing them with unsuitable information (252). Recognizing that an individual may be at different stage of change for different risk factors, each risk factor was individually staged. Thus, for example, a person might be in the preparation stage for physical activity but in the precontemplation stage for smoking cessation.

By 2002, it became obvious that the cardiovascular risk assessment was the most popular with consumers. The decision was made to reduce the number of assessments to one; at the same time, the cardiovascular assessment continued to marketed through two distinctive "brands" or portals (i.e., landing pages): the Heart&Stroke Risk Assessment™ (H&S RA, referred to as the HRA) and the Blood Pressure Action Plan™ (BPAP). Visitors to either site who chose to complete a risk assessment completed the same set of question; the only difference was the branding or name of the assessment.

In March, 2004, the HRA was moved to a new platform that supported the downloading of records into a relational database. The new platform also made it possible to offer an email follow-up service to all users who completed the HRA. Users who enrolled for the service were first asked to select one risk factor on which to focus. Once selected, users were sent a series of emails based on their stage of change as indicated in their HRA. Content of the emails were developed by two clinical psychologists and were designed to move individuals through the stages of change and into making positive behaviour change (e.g., to move from precontemplation to contemplation, preparation and finally action).

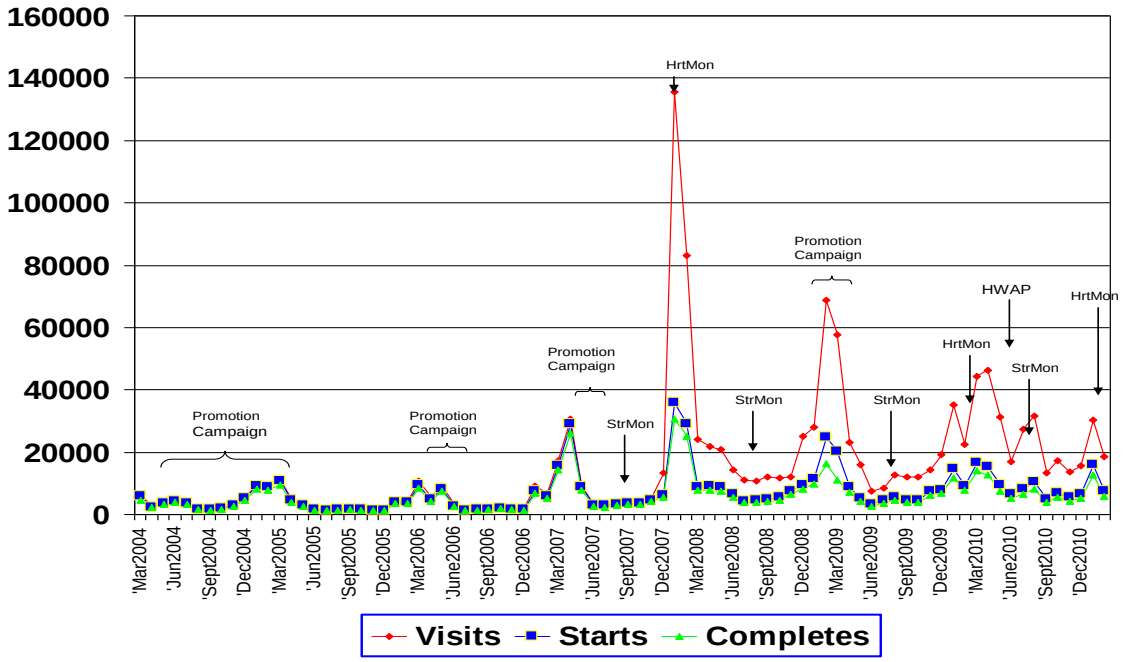
Revisions to the questions asked in the HRA were made in 2006, 2008 and 2010. In December 2010, in recognition of the research potential of the HRA database, a consent question was added, asking if the users would permit their HRA information to be used for research purposes if it was de-identified. As well, staging was changed to include an Action stage (i.e., "I'm already trying to make changes"). In January, 2011, questions concerning socioeconomic status (marital status, highest level of education, employment status, and type of work) were added for the first time.

During most years of operations, the HSFO conducted (with provincial government support) advertising campaigns to drive traffic to the HRA. As well, during several years, the HRA and its website address were included in mass media campaigns released by the organization in support of its Heart Month (February) and Stroke Month (June).

Typically, Heart Month activities were more heavily promoted by the national and provincial organizations and thus received the largest media reaches (typically 50 million or more).

The following figure (Figure 1) shows total number of visits to the HRA landing page, the number of HRAs started and the number of HRAs completed during the first seven years of operation. It shows the strong impact of advertising and promotion on website traffic (e.g., jumps in the number of visits and HRAs started and completed).

Figure 1: Number of health risk assessment visits, starts and completes by month, March 16 2004 to Mar 15 2011



StrMon=Stroke Month media release HrtMon=Heart Month media release HWAP=HWAP media release

As shown in the graph, summer months were typically low-volume periods for the HRA. Traffic begins to increase in January, echoing a pattern reported in a recent study of Canadian Google searches for health information, which found substantive peaks during the month of January (253). However, the biggest traffic gains are associated with promotional campaigns.

In 2011, the HSFO experimented with an incentive-based promotion in collaboration with Air Miles®, a popular credit card loyalty program operated by the company LoyaltyOne, Inc. In late August, 2011, an email blast was sent to Air Miles card holders informing them they could earn ten bonus Air Miles points for completing the HRA and another ten points for enrolling for the email follow-up service (eSupport). As will be described, this promotion resulted in a significant increase in traffic to the eSupport

landing page, the number of HRAs completed, and enrollment for the eSupport follow-up service. The Air Miles promotion can be said to represent a natural experiment of the effects of an incentive on etool uptake and participation.

Table 3: Summary of HRA timelines

Year	HRA Activity
2001	Launch of health etool, consisting of stage-based questionnaires: • A cardiovascular risk assessment • A hypertension risk assessment • A hypertension management assessment
2002	Development of a single cardiovascular risk assessment, marketed through two distinct brands and portals (Heart&Stroke Risk Assessment™ [HRA] and Blood Pressure Action Plan™ [BPAP])
2004	Etool moved to a new portal with capacity to support: • Collection of responses in a relational database • Creation of a stage-based email follow up service
2006	• Minor revisions to HRA (e.g., change in ethnicity question) • Launch of BP self-management module (BPAP)
2008	Minor revisions to HRA
2010	• Major revision to HRA, including addition of questions concerning: • Self-report of non-cardiovascular chronic diseases • Consent for the use of de-identified HRA information for research purposes • Launch of Healthy Weight Action Plan (HWAP)
2011	• January 31, 2011, addition of questions concerning socioeconomic status (marital status, education, employment status, and type of work) • Repackaging of email follow up service as Health eSupport • Launch of Heart&Stroke Risk Assessment™ mobile phone app • August – September, 2011, Air Miles incentive offered for HRA completion and eSupport registration • December 22, 2011, launch of revised version with changes to several questions (e.g. salt questions)

Previous research

Although considerable analysis of the HRA database has been conducted over the years for internal purposes, to date relatively little has been made available to the wider research community. Three publications were released concerning the HRA system in 2011. The first was written for health promotion professionals and looked at the demographic and health profile of HRA users (230), albeit with a smaller and earlier sample of the HRA population (n=45,177, see Appendix 1) then the current study. This article reported substantive and meaningful differences between the general population of Canada and HRA users and discussed the implications for health promotion (230).

The second study was a randomized controlled trial of the effect of the email-based follow up service (eSupport) on hypertension management (254). Of 10,658 users logging onto the HSF website who resided in three recruitment areas in Ontario, 782

(7.3%) completed a telephone screening interview for participation in the I-START (Internet-based Strategic Transdisciplinary Approach to Risk Reduction and Treatment) trial. After application of inclusion and exclusion criteria (e.g., those with cardiovascular or psychiatric diagnoses were excluded, while those participating had to agree to attend pre- and post-treatment clinic visits for biometric measurements), 387 users (3.6% of total or 49.4% of those screened) were included in the study and randomized to the eSupport system or a waiting list control group. The study faced a number of technical difficulties. As eSupport is freely available from the HSF website, in the latter phase of the trial it was found that 35 controls (18%) had accessed the email service despite agreeing to wait until the end of the trial. In addition, only 82 (42%) of experimental subjects met the *a priori* definition of a "therapeutic" dosage of $\geq$8 emails over the four-month study period. These factors were thought to contribute to the lack of effect seen in intent-to-treat analysis. When subjects were divided into groups according to the number of eSupport messages received, those who received what was thought to be the therapeutic dose showed a greater reduction in systolic blood pressure and total cholesterol, but not diastolic blood pressure, compared to the control group that received no messages (254).

The third publication concerned the psychosocial determinants of health behaviour measured during the I-START clinic visits (255). This study found that among the 387 I-START participants, baseline stress and depression were inversely associated with baseline levels of readiness to change exercise and diet behaviours. Receiving the eSupport emails did not appear to change symptoms of psychological distress but compared to controls (no emails) those receiving the therapeutic dose ($\geq$8 emails) showed greater readiness for exercise and diet adherence (255). In summary, the two I-START publications (254, 255) focused upon the efficacy of the eSupport system and were conducted using RCT methodologies. As a result, they are not particularly relevant to understanding the users of a freely-available HRA and do not address the research questions posed in this study.

Creation of the research database

There is currently no single data warehouse for all HRA data points, by which is meant "a copy of transactional data specifically structured for querying and reporting" (256). The research database was created by selecting, downloading and merging two types of information: 1) web metrics captured by the system and 2) users' responses to HRA questions.

1) Web metrics

For the HSF, the vendor operates an online reporting system (ReportWriter) that provides a variety of tables on web metrics, such as number of visitors, number of HRAs started and completed and time to completion. Depending upon the report, information is provided in aggregate by day or for specified time periods. Web metrics also makes it possible to identify records by how they accessed the HRA (i.e., through which portal) and whether they enrolled for follow up etools.

In addition, data on the activities of those who complete the HRA, such as whether they enroll for a follow-up etool, are captured by the system. Such data can be organized by user identification number and transported into Excel spreadsheets.

2) HRA data

Raw HRA data are saved in files that can be downloaded from the ReportWriter site in .cvs format. For the analysis, 23 files were downloaded, extending from January 27, 2010 to December 23, 2011. Each file must be saved in .cvs format and then re-opened and resaved in Excel format before it can be exported into SPSS. Prior to saving in SPSS .sav format, editing is required to ensure that variables are of consistent length and type (the length of some variables changes between different individual files). Records prior to the insertion of the SES question at 21:55, January 31, 2011, and after the uploading of a revised assessment at 1:36 on December 22, 2011, were then deleted.

Height, weight, and actual BMI (as opposed to BMI category) and information on what etools users signed up for are stored in separate files. These files, as well as special data files such as etool engagement, had to be separately downloaded and merged with the main data file.

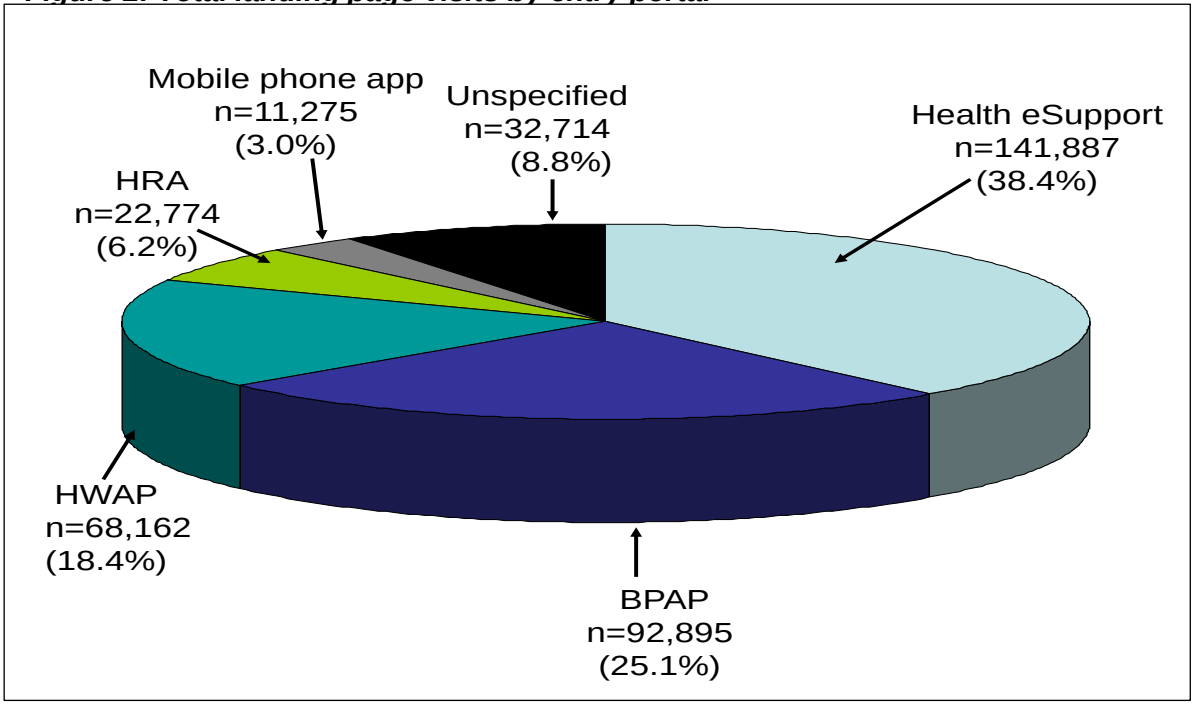
Traffic and uptake

During the period of February 1 to December 21, 2011, there were 369,717 visits to the HRA landing page. Table 4 shows web metrics of how the user accessed the etool (i.e., which portal), the number who started and completed the HRA, and average and medium completion time. Of all records, almost 40% (141,887 or 38.4%) came via the online email coaching portal, eSupport. During the study period, users who came through the eSupport portal constituted the single largest proportion of completed HRAs (86,251 or 57% of all HRAs). Of these 86,251 HRAs, 77,639 (90.0%) were associated with the Air Miles program. Figure 1 is a graphic showing the proportion of total landing page visit by entry portal.

Table 4: Web metrics by portal, February 1 to December 20, 2011

Portal statistics:	BPAP	eSupport	HWAP	Mobile Phone App	HRA	Unspecified	Total
Visits to landing page	92,905	141,887	68,162	11,275	22,774	32,714	369,717
Started HRA (% of visits)	27,938 (30.1%)	106,182 (74.8%)	24,133 (35.4%)	6,685 (59.3%)	13,446 (56.0%)	11,755 (35.9%)	190,139 (51.4%)
Completed HRA (% of started)	21,647 (77.5%)	86,251 (81.2%)	18,846 (78.1%)	4,413 (66.0%)	11,642 (82.1%)	8,829 (75.1%)	151,028 (79.4%)
Avg time to complete HRA in minutes	15.6	19.9	19.1	326.6	34.2	22.0	72.9
Median time to complete HRA in minutes	9.7	11.5	9.4	15.8	9.5	10.1	12.1

Figure 2: Total landing page visits by entry portal



As shown in Table 4, out of 359,373 visits to the etool landing page, in 190,139 cases the visitor started the risk assessment and therefore can be considered a "converted visitor." The overall conversion rate was 51.4%, and ranged from a low of 30.1% for the BPAP to a high of 74.8% for the health eSuppport portal (due, in large part to the Air Miles incentive). In other words, if you exclude the Air Miles incentive, between a half to two-thirds of visits do not result in the start of an HRA. This finding suggests that even at this early stage, there is considerable self-selection among HRA visitors.

In total, 151,028 HRAs were completed, giving an overall completion rate of 79.4%. Completion rates ranged from a low of 66.0% for the mobile phone app to highs of 82.1% for the HRA and 81.2% for the eSupport portal (see Table 4). Unfortunately,

benchmarks for completion rates for freely-available online health risk assessments are not available. It is thought, however, that the approximately 80% completion rate achieved by the HRA may be high. This may reflect a high level of commitment to the process by those who have self-selected to undertake the HRA.

Completing the HRA takes a time commitment of approximately 20 minutes. Table 4 shows the average and median times for completion. Median time may be a more accurate indicator of the time required, as the average is skewed by users who leave the HRA without closing their browser.

For the study period, the relational database contains a total of 147,274 records or 97.5% of HRAs reported to be completed during the study period. A total of 13,754 records are "missing"; these may represent technical problems with the system or cases in which the system was able to identify a duplicate record because the same email address was used for registration.

Of those who completed the HRA 52,915 or a little more than a third registered for one or more of the follow-up etools (see Table 5). Registration was highest for the eSupport email service (41,643, or 28.7% of all new HRAs completed), followed by the HWAP (8,028 registrants or 5.5% of HRAs), and the BP self-management modules (2,287 users or 1.6% of HRAs). The majority of those entering through the eSupport portal were participants in the Air Miles incentive program; if these participants are excluded, the registration rate drops to 5.0%.

Table 5: Registration of new users by portal, February 1 to December 20, 2011

Follow-up etool enrollment	**Portal from which accessed HRA**						
	BPAP n= 21,596	**eSupport n= 86,250**	**HWAP n= 18,793**	**Mobile App n= 4,405**	**HRA n= 10,570**	**Unspecified n=8,809**	**Total n= 150,423**
eSupport (% of completed HRA)	0 (0%)	40,475 (46.9%)	518 (2.8%)	183 (2.8%)	262 (2.5%)	425 (4.8%)	41,643 (27.7%)
BP Module (% of completed HRA)	1,577 (7.3%)	126 (0.1%)	0 (0%)	3 (0.1%)	129 (1.2%)	461 (5.2%)	2,276 (1.5%)
HWAP (% of completed HRA)	1 (0.0%)	359 (0.4%)	6,966 (37.1%)	56 (1.3%)	698 (6.6%)	276 (3.1%)	8,028 (5.3%)
Any online follow up (% of completed HRA)	1,578 (7.3%)	40,960 (47.5%)	7,484 (39.8%)	242 (4.2%)	1,089 (10.3%)	1,162 (13.2%)	52,515 (34.9%)

Figure 3 illustrates the process by which the study database was constructed. Merging resulted in a total of 147,274 records for the study period (February 1 through to

the end of December 20, 2011). Of these, in 141,387 cases (96.0%) users indicated they completed the HRA for themselves, 2,516 (1.7%) for someone else, 3,114 (2.1%) in order to review the site, and in 257 cases (0.2%) this information was missing. Of the 141,387 cases in which users indicated they completed the HRA for themselves, 121,929 (86.2%) gave consent for the use of their information for research purposes, 18,198 (12.9%) denied consent, and a response was missing for 1,260 records (0.9%). Of those who gave consent, 1,412 (1.2%) gave a year of birth that showed age to be <18 years or > 90 years. These records were excluded, as well as 7 records (<0.1%) for which gender was missing. This left 120,510 records, representing:

- 79.8% of all HRAs completed,
- 81.8% of all HRA records saved,
- 85.2% of all assessments created by users for themselves, and
- 98.8% of all assessments for which users gave consent for the use of their information for research purposes.

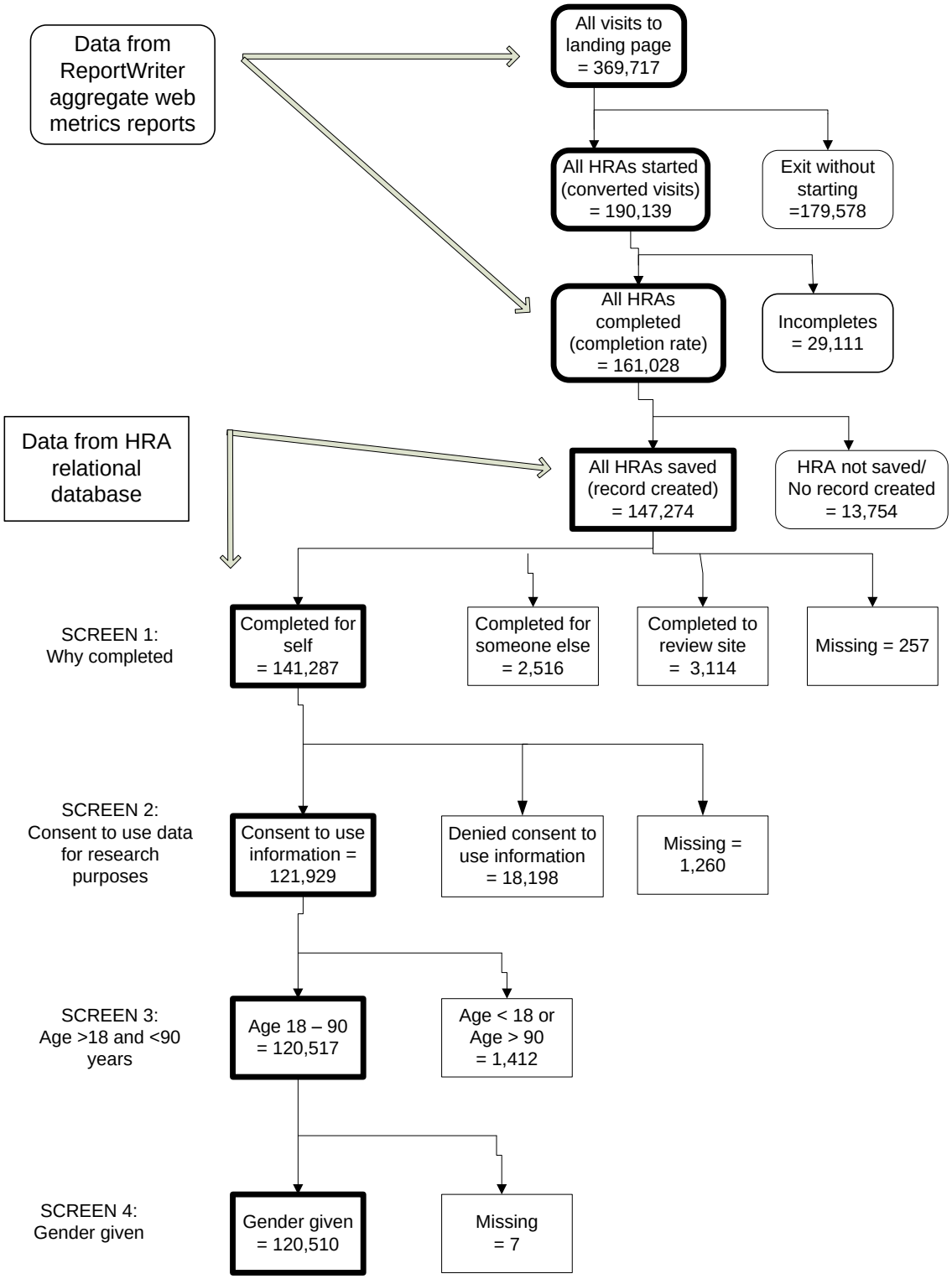
Summary

Freely-available online health risk assessments can attract large numbers of participants, although traffic is not consistent and fluctuates in response to promotional activities. Only a third to half of visits resulted in the user starting the HRA, suggesting considerable self-selection early in the online process. A financial incentive, the Air Miles promotion, increased the proportion of visitors entering through the eSupport portal and of those who completed the HRA (i.e., were converted).

Three-quarters or more of HRA starts resulted in the completion of the assessment. Although benchmarks are difficult to establish, this suggests that those who started the questionnaire were motivated to complete it. However, only relatively small proportions of participants were sufficiently motivated to register for follow-up etools, such as the eSupport email service, the BP self-management module or the HWAP.

Perhaps because the site is sponsored by a recognized health charity, the proportion of users who gave consent for the use of their information for research purposes was high. The data base available for analysis constituted approximately 80% of all HRAs saved during the study period and 85% of all assessments created by users for themselves. As a result, the database created for this study (n=120,510) is probably representative of the larger population of HRA users.

Figure 3: Website traffic, number of HRAs started and completed, and creation of study sample



4. Data and Methods

The HRA questionnaire is divided into six main parts: 1) non-modifiable risk factors, 2) modifiable risk factors, 3) chronic diseases and health care, 4) administrative questions concerning location and how the user accessed the site, 5) marital status, education and employment, and 6) who the user answered the questionnaire for and consent for research. Skip logic is used so only questions appropriate to the user are asked but the overall order of questions has remained consistent since 2004. Please note that a copy of all survey questions and response categories are provided in Appendix 2.

Questionnaire development: wording and order

The HRA etool was not developed for research but rather as a communication and promotion resource for a not-for-profit. As is often the case with not-for-profits, budget and timelines for the development of the original HRA questionnaire were very limited. As a result, there was neither time nor funds for pretesting the wording of questions nor testing for the effect of question sequence.

As described by Wentland and Smith (pg 17-18), response error is associated with three general issues:

1. The respondent does not have access to the information requested (i.e., knowledge);
2. The respondent does not understand the question (i.e., comprehension); and
3. The respondent is not motivated to give accurate information, perhaps due to the sensitivity of the subject (i.e., motivation) (257).

To address these issues in a timely and cost-efficient manner, the following strategies were utilized:

- Knowledge: Respondents were not asked for technical medical information they were unlikely to have ready access to, such as total cholesterol or high density lipoprotein levels or systolic or diastolic blood pressures. Instead, whereas possible questions utilized in the 1998 National Population Health Survey (NPHS) (258) were utilized. For example, the hypertension question asked whether the individual has been diagnosed by a health professional with high blood pressure or has been prescribed medication for the condition. Such information should be known by most respondents.

- Comprehension: Questions were written in as clear a manner as possible and whenever possible were modelled after those in national health surveys such as the NPHS and its predecessor, the CCHS. Such surveys are developed by experts at Health Canada, the national ministry of health, and Statistics Canada, the national agency for statistics. To enhance uniformity of responses and reduce the burden on participants, most questions were closed-ended (259), although a few, such as ethnicity or chronic disease, provided an open-ended "other" option.
- Motivation: Respondents were reminded that their responses were anonymous and confidential. Furthermore, there was also messaging informing participants that the accuracy of their health reports depended upon the truthfulness of their responses.

In addition to review by the team working on the questionnaire development (i.e., the consultant who wrote the questions, the program manager, and the developers), questions and responses were reviewed on a *pro bono* basis by two clinical psychologists with an interest in the Transtheoretical Model of Change, one university- and the other practice-based.

Between 2004 and 2013, in order to facilitate longitudinal analysis of records, there was a deliberate policy of keeping changes to the wording of questions to a minimum. Exceptions were:

- the ethnicity question, which evolved from a simple dichotomous response option ("Are you South Asian [India, Pakistan, Sir Lank, Bangladesh], First Nations/Aboriginal, Inuit or Black?") to four and eventually 13 response options
- age, which was originally asked as a categorical question; from 2009 and throughout the data collection period, age was asked in the form of year of birth
- the list of "how you heard of the website" options (these tended to change as marketing approaches changed)
- location, which changed from a list of provinces to the addition, in 2009, of a request for the first three digits of the postal code (i.e., forward sortation area)

Even when the etool was "reskinned" (i.e., the look redesigned), wording was kept as consistent as possible. In 2009, a plain language consultant was contracted to do a review of the questions and as a result some descriptions of chronic conditions were

revised. There was no change to the wording of any question throughout the study period.

As noted by Rea and Parker (1997, p. 35), the order of question can affect responses, as a “poorly organized questionnaire can confused respondents” and “bias their responses” (259). Common strategies to address this issue have included grouping related questions together, asking sensitive questions later in the questionnaire, when the respondent is presumably more comfortable with the process, and putting questions in a logical sequence that facilitates memory retrieval (259). However, the effect of question order may depend upon what the researcher is attempting to study. One study that included two multi-question modules – one on overall health status and the other on symptoms specific to a specific disease – found changing the order was not associated with any significant differences in scores (260). In other words, general health questions may or may not be needed to “warm up” respondents to a disease survey. Another study, concerning the reason for Emergency Department visits in Australia, found that when the order of questions was randomized more reasons were selected, compared to when the order was fixed with the most common conditions being given first (261). Based on these findings, the author suggests that batteries or long lists of questions should, whenever possible, be randomized (261). Such an approach might be feasible for a small number of questions in the HRA, such as the list of chronic conditions or ethnic groups, although further research would be needed to determine if the benefits would justify the additional programming, design and data management costs.

The basic sequence of questions in the HRA was established in 2000 and utilized the principles of grouping related questions and positioning more sensitive questions later in the questionnaire (259, 262, 263). Questions were grouped into four main categories: 1) non-modifiable risk factors or those factors such as family history over which the respondent has no control, 2) behaviourial, modifiable risk factors such as physical activity, diet, stress, smoking, diet and alcohol consumption, 3) the presence or absence of chronic disease and screening, testing and management of hypertension, dyslipidemia and diabetes, and 4) non-medical questions such as source of primary care, education, occupation, marital status, location and how the person heard of the etool.

Non-modifiable questions were positioned first, in hopes that people would find these more nonthreatening than questions about their personal health behaviours. Each section was preceded by a short explanation (one or two sentences) explaining why these sorts of questions were important.

The basic sequence of questions did not change over time. When new questions were added, they were always added after existing questions. In 2008, for example, the

list of chronic conditions was added after the existing questions on modifiable risk factors; the list reflected the most common chronic diseases reported by the Ontario Ministry of Health and Long-term Care. In 2009, questions on source of primary care, marital status, education and occupation were added and these too were added after all existing questions.

Skip logic was used so users were not asked inappropriate or irrelevant questions (e.g., a person who reported being physically active would not be asked when he/she might be willing to start becoming more active). Although users could use a back button to change previous answers, all health questions were mandatory and each screen (page) of questions had to be completed before the user could proceed to the next. Only the questions on primary care, marital status, education, occupation, location, and how the user heard about the site were optional; however, as there was no notice that these questions were optional some users may have assumed they were also mandatory.

There were no changes in the wording or sequence of questions throughout the data collection period, nor to the site design.

Non-modifiable risk factors

Users were asked their gender, year of birth (from which age was calculated), and ethnicity. Family history was defined as having blood relatives ("your natural or biological parents, grandparents, brothers, sisters, or children") with a history of premature heart disease (a female relative prior to age 65 and/or a male relative before age 55), premature stroke (prior to age 65), hypertension or dyslipidemia (the latter defined as high cholesterol, hypercholesterolemia, an unhealthy cholesterol profile or high triglycerides). Response options were "yes," "no" and "don't know." For analysis as binary variables, responses were recoded into family history present ("yes" responses) or family history absent ("no" and "don't know" responses).

For ethnicity, users were asked to choose the one ethnic group to which they most see themselves as belonging (i.e., strongest identity). Categories commonly used in the Canadian Census were used but unlike the Census, multiple responses were not accommodated. Three categories were identified by the HSF as being at increased risk for CVD: South Asian (described as including Indian, Pakistani, Sri Lankan, Bangladeshi, etc.), Aboriginal North American (First Nations, Inuit or Métis), or of African descent (i.e., Black). Belonging to one of these groups was coded as being positive for having the non-modifiable risk factor of high risk ethnicity.

For analysis, a count of total number of modifiable risk factors was created, with a range of zero to six. The count consisted of number of diseases for which the user reported a family history, as well as self-identification as belonging to one of the three increased-risk ethnic groups. Although age is a non-modifiable risk factor for CVD it was not included in the count because it was frequently used as an analytical variable. For analysis, age was used as an interval variable or combined into groups.

Modifiable risk factors

The HRA addresses seven modifiable risk factors for CVD:

1. Physical activity: This user is asked about moderate activity at work or at home, defined as activities such as brisk walking, active gardening, swimming, dancing or biking, for at least 30 to 60 minutes four or more days of the week. Unlike the CCHS, the question is not limited to leisure-time activity (264).
2. Smoking: The user is asked if they smoke but there are no follow-up questions to verify amount or to distinguish between never and former smokers.
3. Being overweight: Users are asked to enter their height and weight, from which the system calculates their body mass index (BMI). Overweight or obese is defined as a BMI $\geq$25 kg/m^2. In addition, the user is asked to enter the waist measurement, from which is calculated the risk category according to the classification system in the Canadian Obesity Guidelines.
4. Higher salt consumption: The user is asked to indicate whether high-salt foods are frequently eaten, whether he/she tries to monitor salt intake, or makes a strong effort to limit salt intake. Based on the understanding that the average Canadian diet is relatively high in salt (265), the user is identified as having high salt consumption if responses indicate high-salt foods are frequently eaten or salt intake is not controlled or monitored.
5. Alcohol consumption: Users are asked whether their alcohol consumption exceeds the Low-Risk Drinking Guidelines daily and weekly maximums by gender as posted in 2011 (266). To help the user respond, the definition of what constitutes one drink is provided.
6. Dietary behaviours: The user is asked how frequently (less than once a week, 1-2 times a week, or 3 or more times a week) high fat foods, fast foods, foods rich in omega-3 oils such as cold water fish, and five or more servings of vegetables and fruit are consumed. For each, definitions are given of what types of foods are defined and, for vegetables and fruit, what constitutes a serving. For analysis, poor dietary behaviour is defined as eating high fat foods or fast foods 3 or more times a week,

eating fish less than 1-2 times a week, or eating five or more servings of vegetables and fruit less than three times a week.

7. Stress: Users are asked how frequently in a typical week they feel overwhelmed or stressed by the demands on them. Response options are “seldom or never,” “a few times,” and “often or most of the time.” Those indicating “often or most of the time” were coded as positive for frequent stress.

For each negative behaviour, the user is asked to indicate their stage of change for that behaviour utilizing cut-points commonly used in studies involving the Transtheoretical Model of Change (251): already started or working on it (representing the Action stage), within the next 30 days (Preparation stage), within the next 6 months (Contemplation stage), or not planning to change the behaviour (Precontemplation stage). For physical activity, alcohol and smoking, response options incorporate the stage of change whereas for salt and alcohol consumption, stress, being overweight, and $\geq$1 poor dietary behaviours, stage of change is asked as a follow-up question to those who report a risk behaviour. Although these cut-points are commonly used, it should be noted that they were developed primarily in smoking cessation research and questions have been raised as to whether they are valid for other types of behaviour, such as diet (267).

For analysis, modifiable risk factors were utilized in the following manner:

- As binary variables (risk factor present vs. risk factor absent)
- Count or sum of the number of risk factors coded as present, ranging from zero to seven
- Each of the seven behaviours was re-coded as a 5-point nominal variable incorporating the stage of change, with a higher score indicating healthier behaviour or greater readiness to change. For example, for physical activity being active would be scored as five, inactive but in the Action stage as four, inactive and in the Preparation stage as three, inactive and in the Contemplation stage as two, and inactive and in the Precontemplation stage as one.
- A Lifestyle Healthiness Score was created by summing the staged score for the seven modifiable risk factors. For this variable scores could range from seven (unhealthy and/or low willingness to change) to a maximum of 40 (healthy lifestyle with no modifiable risk factors).

Chronic diseases

Users were presented with a list of 16 common adult chronic diseases and asked to indicate if "a doctor or other healthcare professional ever told you that you have any of the following chronic (long-term) conditions." Wording differs from that in the CCHS in that the CCHS specifies the condition is "expected to last or have already lasted 6 months or more and that have been diagnosed by a health professional" (264). Other chronic conditions could be entered into a text box. Users were also asked to click if they had been prescribed medication for a condition by their healthcare providers. Those who indicated $\geq$1 condition for which they were prescribed medication were asked a follow-up question on overall medication compliance: whether they miss taking their medication as prescribed most of the time, some of the time, seldom or rarely, or never.

Two of the conditions on the list constitute CVD: 1) heart attack or heart disease and 2) stroke or "mini-stroke" (transient ischemic attack or TIA). Three conditions in the list are considered risk factors for CVD: hypertension (high blood pressure), dyslipidemia (explained as in the family history question) and diabetes (type 1 or 2). Information from the list of chronic diseases were utilized in analysis in the following ways:

- As binary variables (condition present or absent)
- Count of the total number of vascular diseases, with "vascular disease" defined as the two CVD conditions (heart disease, stroke/TIA) and the three conditions that are proven risk factors for CVD (diabetes, dyslipidemia and hypertension).

For hypertension, diabetes, and dyslipidemia those not reporting the condition were asked a follow-up question to capture information on preventative screening. Those who reported one or more of these three conditions were asked two follow-up questions. These questions addressed 1) interval of time since last tested by a healthcare provider (e.g., for those with diabetes, time since last hemoglobin A1c test), and 2) self-report of how frequently the condition indicator (i.e., blood glucose, blood pressure, or lipids) is in what is considered a "healthy range or in the range recommended by your healthcare provider." Response options for this question included most of the time, some of the time, seldom or rarely, never, or don't know. For analysis, any response other than "most of the time" was considered an indicator of sub-optimal condition control.

In addition, users were asked if they have a healthcare professional they consider to be their family doctor or primary healthcare provider, as well as where they go for most of their medical care (physician's office, walk-in clinic, hospital emergency department, or other).

Derived variable

For analysis, the following variable was created:

- Total number of health concerns: A count of the total number of non-modifiable and modifiable risk factors and vascular diseases reported by the user. This variable could range from zero to 18 (6 non-modifiable, 7 modifiable, and 5 vascular diseases).

Administrative questions

For administrative purposes, two questions were asked. First, users were asked to give the first three digits of their postal code (the Forward Sortation Area code) or, alternatively, to indicate their province of residence or residency outside of Canada. A question also asked users how they learned about the web site (e.g., brochure or poster, online advertisement, etc.). The latter question was not utilized for the current analysis.

Marital status, education and employment

Prior to the study period, the HRA had contained no questions regarding marital status, education and employment. The questions that were added were based largely on those utilized in the CCHS. It was hoped that education could be used as a proxy for socioeconomic status (SES), as the HSF did not want to query users on income.

Education was captured in five categories similar to those used in the CCHS: less than high school, high school, some post-secondary education, college or university graduate, or rather not say. For some analyses, education was recoded into two groups: less education (less than high school or high school) *vs.* more education (some post-secondary or college/university graduate).

Who completed assessment for and consent

The last two questions asked the user for whom the questionnaire was completed: for self, someone else, or to investigate or review the site. If the person indicated they completed the assessment for him/herself, the consent question was asked. The question asked whether de-identified information could be included in an anonymous research database.

Engagement data

Separate data files provided by the vendor provided the following data points:

- Landing page or portal from which the user accessed the HRA (HRA landing page, mobile phone app, BPAP landing page, eSupport landing page, or HWAP)
- Whether the user came through the Air Miles promotion
- Whether after completion of the HRA the user enrolled for eSupport, the BPAP self-management module or the HWAP
- Number of times users who enrolled for eSupport interacted with the system (i.e., choose a risk factor to focus on or rescored readiness to change)
- For those who rescored their readiness to change in the eSupport system, revised readiness to change stage.

Using unique case record identification record, these additional data points could be merged with the main HRA data base.

Validity of self-reported data

One of the challenges of the HRA data is determining the quality and validity of the data. External validity is a major concern as it determines the extent to which results can be generalized to others (268). For an observational data base such as the HRA, there are also concerns about internal validity, i.e., the extent to which questions accurately capture what they are supposed to measure (269). Since the questionnaire is designed as essentially a one-time event, the issue of reliability (the extent to which measure are replicable over time) (269) is still relevant but may be less pressing.

As assessments are completed remotely and anonymously, there is no means of validating responses. Some research suggests that self-reported health data may underestimate the proportion of individuals “at risk” or with health risk factors (270). However, this may vary between users and according to the type of behaviour being queried.

Validity of the self-report of two of the modifiable risk factors in the HRA, smoking and BMI (as estimated from self-reported height and weight) has been studied. One Canadian study found self-report of smoking status has a sensitivity of more than 90% (271), while an American study determined the prevalence of smoking in self-reported online panels was comparable to that obtained through national representative surveys

(272). In other Canadian analysis, sensitivity for BMI from self-reported height and weight compared to those obtained from measurements was 58.5% for males and 68.5% for females, with specificities of, respectively, 98.4% and 99.2% (273). Other research has reported under-reporting of weight and over-reporting of height in self-reported Canadian Community Health Survey (CCHS) data, with the magnitude varying by gender, age, and BMI category (274).

HRA questions on diet, stress, physical activity and salt consumption were not based on existing measures used by surveys such as the NPHS or CCHS. It should be noted that in 2014 the HRA is being revised so questions on physical activity and diet will reflect those in the CCHS.

There is some evidence of fair to good validity for self-report for health conditions and medical conditions from Australia (275) and the U.S. (276). Although a recent meta-analysis questions the accuracy of self-reported hypertension (277), in one American study the sensitivity for self-report of hypertension was 83% (specificity 81%) and for diabetes 73% (specificity 99%) (276). In contrast, a 2012 study in the Netherlands found the sensitivity and specificity to be 38.9% and 98.0% for hypertension, 76.8% and 98.8% for diabetes, and 80.9% and 75.7% for overweight (278). In another analysis of American data, prevalence estimates of hypertension were found to be similar to examination-based estimates but self-report of hypercholesterolemia significantly lower (279). It is possible that differences may represent situations in which conditions have not been diagnosed, rather than inaccurate responses of users. In a 2008 survey of Ontarians, for example, 13.7% of those with hypertension were unaware of their condition (280).

Some research has been conducted concerning the validity of self-report for family history of cardiovascular conditions. One study in the U.S. reported the sensitivity of report of a family history of coronary heart disease was 87% for spouses, 85% for parents, and 81% for sibling (281). For diabetes, the numbers were, respectively, 83%, 87% and 72% and for hypertension 77%, 76% and 56%. In this study, specificity values were above 90% for most comparisons (281). Age, gender, disease status and ethnicity tended to influence the accuracy of reported sibling disease history but had little effect on spousal or parental medical history (281). In a more recent study from the Netherlands, when self-reports were compared to the parents' or siblings' own self-report, sensitivity and specificity were 89.2% and 81.0% for diabetes and 92.2% and 56.2% for hypertension (278) However, sensitivity and specificity were lower when reports were compared to physician-assessed health status of relatives: respectively, 70.8% and 77.8% for diabetes and 67.4% and 63.2% for hypertension (278). In other words, self-

report of family history of cardiovascular-related conditions in the HRA may be capturing *perceived* family medical history and be less accurate for capturing actual history.

In summary, although the validity of HRA self-reports cannot be established, review of the literature suggests that for most data points sensitivity may be at least 60% and specificity almost as high. It could be argued that as the HRA is anonymous and completed at the time and place of the user's choosing, responses may be less influenced by the desire to give an interviewer "socially acceptable responses" (282), as well as acquiescence effect or the tendency to provide affirmative answers (283). At least one study has suggested that completion and accuracy of web-based surveys may be better than telephone-based questionnaires (284).

Socioeconomic status indicators

The HSF did not want to ask income so highest level of education and type of work are the only available indicators of SES. This is unfortunate, as composite measures of SES that include multiple measures such as area and household income and education may be preferable (285). The relationship between education and health may be complex, as education correlates with income, employment, place of residence and health literacy (286). However, education may be useful in the study of health. There is Canadian research showing a negative relationship between education and all-cause (287) and cause-specific mortality (288) and the use of medical (289) and mental health services (290). There is also evidence from Canada that education has a negative relationship with the risk of cardiovascular disease (291), diabetes (292), and Alzheimer's disease (293). As well, education has been estimated to be responsible for 24% of the population attributable risk for lung cancer in males and 19% in females (294). The relationship between education and health outcomes is strongly influenced by education-related gradients in behavioural risk factors such as smoking, physical inactivity, being overweight, and meeting the daily recommended servings of vegetables and fruit, although there may be some divergence by gender (295, 296).

In short, although for research purposes information on income might be optimal, there is evidence suggesting education may be a useful proxy for socioeconomic status.

Methods

Method of analysis varied according to the research question being addressed.

Descriptives (Chapter 5)

With the exception of age, most of the data points created by the HRA were nominal. For nominal data, the primary descriptive statistic was proportions, although counts were also created (e.g., number of chronic diseases reported) that could be presented as means and medians. Proportions and means were given by common demographic variables such as gender, age, and level of education.

In determining whether differences between groups are meaningful, the large size of the data base posed a challenge for inferential statistics. As succinctly noted by Rex Kline (297), "If you increase the sample size enough, any result will be statistically significant" (pg. 16) – even though it may not be important or "clinically significant" (298), i.e., meaningful for the program or intervention. In fact, as shown in the tables in the Appendices, because of the large sample size almost all comparisons were statistically significant at the $p<.001$ level. In the text, therefore, $p<.001$ was not reported; instead the relatively less common occurrences in which p was >.001 were noted.

Given the limitations of inferential statistics, how can it be determined if a difference between groups is meaningful? One option might be to establish a minimal difference required to be considered meaningful (e.g., a relative difference of, for example, 5% or 10%). The magnitude of the difference would, however, be arbitrary, particularly in light of the lack of similar analyses of freely-available risk assessment data. Another option, and the one adopted in this study, was to use effect size as a measure of the magnitude of the difference between groups (297, 299). Effect size estimates utilized in this study were:

- For comparing the means of two groups, Cohen's *d* index was calculated using the University of Colorado Colorado Sprint (UCCS) online effect size calculator for two independent populations (*www.uccs.edu/~faculty/lbecker/*, accessed 7/05/2013). A hand calculation of the Cohen's *d* value for a sample was conducted and results were compared to another online calculator (*www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html*, accessed 7/05/2013) to confirm that the online calculator was accurate. According to Sheskin (2007, pg 169) standard practice is to consider a Cohen's *d* of 0.2 a small effect, 0.5 a medium-sized effect and 0.8 or greater a large effect (300).

- For means for more than two groups, the preferred measure of effect size was omega squared (ϖ). Omega values were calculated from sum of squares, degrees of freedom and mean square values from ANOVA tables in the manner described by Field (pg. 389) (301) . According to Sheskin (2007, pg 449-450), a small effect is indicated by a

value great than .0099 (i.e., .01) but not exceeding 0.588, a medium-sized effect between 0.588 and 0.1379 (i.e., 0.60), and a large effect greater than .1397 or approximately 0.14 (300).

- For categorical data, the main measure of effect size was Cramer's *phi* coefficient (Cramer's V), a measure of the relative strength of the association between two variables. Cramer's V ranges from 0 (no relationship) to a maximum of 1.0 (perfect relationship). It is a preferred effect when a table is larger than 2 x 2, as it can take into account the degrees of freedom (302). As noted in Crewson's *Applied Statistics Handbook*, Cramer's V is particularly useful in situations where statistical significance of a chi square may be unduly influenced by a large sample size (303). As reported by Pallant (pg. 217) the general rule of thumb for interpreting Cramer's V are: for two categories (1 degree of freedom) .01 represents a small effect, .30 a medium effect and .50 a large effect and for three categories or 2 degrees of freedom (either number of rows or number of columns minus one is equal to 2), .07 represents a small effect, .21 a medium effect and .35 a large effect (302).

- As described by Sheskin (pg 130) when interval data are used or implied for at least one categorical variable, it may be appropriate to report the eta squared (*eta*) statistic (300). Age group was initially categorized into five groups (18-34, 35-44, 45-54, 55-64, 65-74 and 75-90) and level of education into four (no secondary school, secondary school, incomplete college or university, and completed college or university). Such variables can be said to approximate interval data, in that they represent ordered or progressive increases (e.g., category 2 is older or more highly educated than category 1, etc.), even though categories may not be equal in size (304). Thus, *eta* was reported when age or level of education were cross-tabulated with other categorical data. *Eta* is an estimator of the strength of the association between variables and thus ranges on a scale from zero (no association) to one (maximum association) (299). As described by Pallant (240), a general rule of thumb proposed by Cohen in 1988 is that .01 represents a small effect, .06 a moderate effect and .14 a large effect (302). Although *eta* tends to overestimate the level of association, Grissom and Kim (pg. 12) report the bias is reduced when sample sizes are larger (299).

Comparisons (Chapter 6)

Three comparisons were undertaken in this chapter:

1) The HRA to the general population of Canada;
2) To determine the effect of an incentive, a comparison was made between HRA users brought in by the Air Miles promotion to those who were not

3) To evaluate the generalizability of samples created for RCTs, a comparison of the HRA to three RCT samples

1) The HRA to the general population of Canada

Although it may be tempting to assume that HRA users will not be representative of the general population because they are health information seekers, testing is required to determine if this is the case. For example, it could be argued the HRA sample may be generalizable because of the high prevalence of CVD risk factors: a recent study has estimated that nine out of ten Canadian adults have one or more of six major CVD risk factors (smoking, physical inactivity, overweight, poor diet, diabetes or hypertension) (305). A comparison of Canadian HRA users to the general population of Canada is needed to objectively determine the issue of generalizability.

To compare age and gender, numbers of Canadians in the general population by age and gender were downloaded from Statistics Canada national census files (306). Because the 2011 Canadian census had a global non-response rate for level of education of 26%, education and estimates of health behaviours were derived from the self-reported 2010 Canadian Community Health Survey (CCHS), as available from Statistics Canada' CANSIM system (307). The CCHS is a cross-sectional, national survey conducted by the federal government that collects information on health status, health care utilization and health determinants from a representative sample of approximately 65,000 Canadians aged 12 and over, excluding institutional residents, full-time members of the Canadian Forces, and residents of certain remote regions. Estimates were rounded and only those based on a sample greater than 30 (i.e., with a coefficient of variation less than 33.3%) were cited.

Of 120,510 records in the HRA, 1,702 (1.4%) users indicated that they did not live in Canada and so were excluded. For comparison to the Census Data and the CCHS, 1,146 (1.0%) records for which the age was given as less than 20 years were excluded, leaving 117,690 records. This represented 97.7% of the original HRA research database. To explore whether differences between the general population and the HRA sample may be due to gender or age, highest level of education will be compared using both unadjusted and adjusted (i.e., weighted) proportions.

2) Comparison of Air Miles and non-Air Miles HRA users

Analysis was conducted to determine if the use of an incentive (which is used in some RCTs) influenced the type of people who complete a cardiovascular health risk

assessment. This analysis took advantage of the natural experimental provided by the HSF's Air Miles promotion.

3) Comparison of the HRA to three RCT samples

For this analysis, the literature was scanned for etools similar to the HRA and three studies were identified. The first, by Wanner *et al.* was of a physical activity etool marketed to the Swiss general population and included both an open-access and closed-access, RCT arm (308). The second described an online CVD risk assessment tested among employees in the Netherlands (309). The third concerned a Dutch web-based tailored lifestyle intervention addressing many of the behaviours highlighted in the HRA, such as physical activity, diet, alcohol consumption, and smoking (310).

Methods

For comparisons 1 and 3, the main measure of effect size was the odds ratio (OR), a measure of "how many times greater the odds are that a member of a certain population will fall into a certain category than the odds are that a member of another population will fall into that category" (pg. 188) (299). The OR is appropriate as it can be calculated for cross-sectional, point-prevalence data such as the HRA and CCHS (311). The value of an OR can range from zero to infinity and be either negative (reduced odds) or positive (increased odds). ORs and their associated 95% confidence intervals were generated by using the 2 x 2 odds ratio calculator from Vassar College (US) (*http://faculty.vassar.edu/lowry/odds2x2.html*, accessed 7/05/2013). A hand calculation of an OR showed the online calculator was accurate.

As described by Olivier (2013), various cut-points have been suggested for interpreting ORs, depending in large part on assumptions about probabilities within the sample (312). For example, Cohen recommended 1.49, 3.45 and 9.0 as indicators of small, medium and large effects, whereas Ferguson (313) suggested 2.0, 3.0 and 4.0, and Olivieri, excluding assumptions as to marginal probabilities, 1.22, 1.86 and 3.00 (312). Based on several previous reviews, Olivier suggests that in the social sciences, an OR of 2.0 should be considered the recommended minimum effect size (RMPE) to identify a "practically" significant effect (313). Using the RMPE will help to ensure that effects are not exaggerated, as the OR may over-estimate the likelihood of an outcome when it is common in both groups (311). In interpreting the difference between adjusted and unadjusted ORs, the cut-off point of a 10% relative difference will be used to indicate a situation in which the weighting variables are confounders, as suggested by Hernan *et al.* (314).

Segmentation (Chapter 7)

As described by Bailey, classifying objects into groups "is arguably one of the most central and generic of all our conceptual exercise" (315). Categorizing people into groups is the starting point for tailoring and targeting, which are considered major factors influencing health promotion efficacy (10, 168, 316). The two terms are not synonymous. Targeting refers to customizing by demographic categories, such as gender, ethnicity or age (317, 318) and is based on the assumption that all people in the same demographic group have the same information needs. Tailoring, on the other hand, is based on segmenting the audience into groups based on needs, attitudes or behaviours (317). Such segments may cross demographic categories.

As health information seekers comprise a substantive proportion of the population, perhaps 60% or more (51), it is unlikely they form a single, homogeneous group. As suggested by the literature (317, 318) standard demographic categories may be of only limited utility in creating clearly-defined groupings of health information seekers. Alternative statistical approaches may be needed in order to understand and differentiate the HRA population, such as segmentation.

The use of segmentation to analyze and segment the audience of health promotion programs has been used in social marketing to ensure a better understanding of users and therefore an enhanced consumer orientation (319). Segmentation for social marketing began in the 1980s and was refined during the 1990s (320, 321), leading to what Noar *et al.* (2007) refer to as a "blossoming literature on tailored communications" (322). As noted by Noar *et al.* (322), whereas targeting addresses groups, typically based on demographics such as age, gender or ethnicity, tailoring adapts messaging to individuals based on characteristics that may transcend demographic categories. These characteristics may reflect needs or preferences (317), behaviours (323), a combination of communication, behavioural, psychological or demographic dimensions (324), or health-related constructs such as health self-efficacy, health information seeking behaviours and attitudes, prevention orientation, relationship with health care providers (325). Although demographics may influence health consumer segments, they do not define them; segments typically transcend demographic strata and different demographic groups may have different segment profiles (326).

Noar *et al.'s* 2007 meta-analysis comparing generic to tailored print health resources found that tailoring increased the effectiveness of health promotion and behaviour change messaging (322); an earlier, non-systematic review also reported that, compared to generic messages, those that were tailored were better remembered, more

likely to be read and perceived as more relevant or credible (327). It has been suggested that the personal relevance introduced by tailoring reduces the tendency of subjects to rely on heuristic "short-cuts" and, as a result, to consider the information more carefully (i.e., in the elaboration likelihood model, to engage in central, as opposed to peripheral, route processing) (322). Rimer and Kreuter (2006) suggest tailoring enhances motivation through multiple pathways, such as identifying design and production elements that are more likely to capture the individual's attention, better matching the amount, type, and delivery channel of the content to his/her needs or interests (which, as discussed, may transcend demographic groupings), and framing the information in a meaningful context (328). By doing so, tailoring may increase not only information reception or attention, but facilitate acceptance and utilization (referred to as "yielding") (328). Compared to print materials, web- and computer-based resources have even greater ability to generate tailored information matching the needs of individuals; they also possess the capacity to provide personalization (e.g., use of the individual's name, although this approach has lost its novelty and is losing credibility) and feedback (329).

The utility of segmentation has become so accepted that it is almost ubiquitous in some health promotion/social marketing (330). A number of different procedures can be used to create segments.

Cluster analysis is a generic term referring to a number of mathematical procedures to group data into sets (331, 332). Cluster analysis can be used in an exploratory manner (i.e., to create a question or hypothesis) or to test a hypothesis (i.e., to confirm or disprove a grouping obtained in some other manner) (331). Romesburg refers to exploratory cluster analysis as a form of retroduction or the development of a hypothetical reason (or research hypothesis) based on observations from observed facts (331).

As discussed by Dolnicar (2005), one of the most common misperception about segmentation is the assumption that groups are always naturally-occurring and are clearly distinct entities (333). In reality, segments are artificially-constructed groups with often indistinct boundaries (333). Moreover, different types of cluster analysis are based upon different procedures or criterion for creating groups (332, 334) and thus can result in the generation of dissimilar groupings. Even if no natural structure exists in a data set, Dolnicar argues that segmenting is still beneficial as it eliminates the simplistic and often misleading assumption that a population is monolithic (333).

For this study, two forms of cluster analysis that are included in the standard SPSS package were utilized. They were:

1. K-means: a form of partitional clustering that is capable of handling large data bases (335). This procedure is referred to as a relocation method as cases are classified and reclassified until cluster means stop changing significantly; at this point, the means of clusters are calculated a last time and group membership is assigned (336). Advantages of the k-means procedure for analyzing data sets generated by freely-available health etools include its ability to handle large databases and the fact that it is available in a number of standard statistical packages (e.g., SPSS, SAS, Systat) as well as some freeware programs (R, ELKI, Weka, etc.). The limitations of the k-means include its inability to handle categorical data and the fact that the user must specify the number of groups.

2. Two-step: SPSS offers a clustering procedure that drives its name from the fact that it consists of two calculations (337). In the first step, sequential clustering is used to form groups into a modified cluster tree and Bayesian Information Criterion (BIC) values are calculated to find the initial estimate for the number of clusters. The second step refines the initial estimate by identifying the greatest change in distance between the closest clusters and using this information to combine groups into the desired number of clusters. The SPSS two-step procedure generates an estimate of the extent to which groups are internally cohesive and separated from one another, referred to as the silhouette co-efficient (337). Advantages of the two-step procedure are that it can handle both categorical and continuous variables, the number of groups can be generated by the system or specified by the user, it can handle large data bases, and it produces an estimate of the strength of the solution. Perhaps the most important disadvantage of the procedure is the fact that it is not provided within the suite of clustering procedures offered by other common statistical packages, such as SAS or Systat. However, it should be noted that two-step clustering can be conducted using other statistical packages by first calculating and saving the distance between data points and subsequently submitting this data set to hierarchical clustering (337).

SPSS also provides a hierarchical clustering procedure. This procedure is suitable for smaller data bases (336) and is computationally too demanding for efficient analysis of a set of over 120,000 records.

In addition to the two forms of cluster analysis, latent class analysis (LCA) was conducted. Unlike cluster analysis, which is primarily oriented towards the production of homogeneous groups, LCA assumes variables are independent of one another (338) and groups are formed on the basis of the relationship of the clustering variables to an unseen (i.e., latent) variable (332, 339). Exploratory LCA is often used when there are several measures which are thought to be parts of a common complex (339), such as when

different attitudes or behaviours are thought to be part of an overarching construct (31). For example, LCA may be helpful if the different health-related behaviours captured in the HRA are related to a common construct such as being health conscious or health conscientious. Over the past decade, with the increasing availability of commercial or freeware statistical packages that can conduct LCA, this form of analysis has been increasingly used by polling firms in order to segment populations. For this study, the commercial statistical package LatentGold® ™ 4.5. by Jay Magidson and Joeroen Vermunt of Statistical Innovations Inc. (339) was utilized.

One of the greatest challenges in exploratory cluster analysis is to set the number of groups or clusters. The number of clusters may be based on similar, previous analyses, "expert opinion," or, in the case of hierarchical clustering, making a subjective decision as to where a natural "cut" in the data occurs (332). For those working in real-world applications, it is typical to look for segments that are large enough to be practical for programming or marketing efforts (340). For example, an analysis of 243 studies found most (two-thirds) preferred solutions of between three to five clusters, with 23% three clusters, 22% four and 19% five (341).

For this study, because the goal was to conduct analysis that may have practical applications, a decision was made to focus on four-group solutions. This decision was arbitrary but reflects the number of groups for which it would be feasible to develop tailored resources. However, other solutions (e.g., three and/or five-group solutions) were also tested to see how the data would react.

If possible, the selection of clustering variables should be driven by the research hypothesis or understanding of the population. Mooi and Sarstedt recommend that the number of variables for clustering be kept to a minimum as using too many clustering variables increase the odds of high collinearity; this in turn may cause over-representation of the shared factor(s) and reduce the procedure's ability to identify distinct segments (340). However, what constitutes "high collinearity" has not been standardized. Sambandam, for example, suggests that clustering should not be conducted if the correlation coefficient between the variables exceeds .500 (342) , whereas Mooi and Sarstedt set the bar much higher, at values exceeding .900 (340).

Validating segmentations

Validating a solution is a challenge in exploratory segmentation. As noted by Stockburger, "Cluster analysis will always produce a grouping" (343). In other words,

software packages can easily spit out some solution: the challenge is to produce a cluster solution that is robust and appears to have validity (331). A number of methods have been suggested for validating segmentations, ranging from comparing clusters to known groups or the opinion of experts, to replication (e.g., clustering on a split sample of the data) or using multivariate techniques such as discriminant analysis (344) or multivariate analysis of variance (MANOVA) (331). Aldenderfer and Blashfield argue that replication is a check on internal consistency but does not mean the solution has external validity (332). They also argue that although discriminant or other multivariate techniques based on the clustering variables are frequently used to test solutions, they are inappropriate since they will report significant findings even if in reality there are no clusters in the data (332). Aldenderfer and Blashfield's preferred method of validating a solution is to compare groups on variables not used to generate the clusters (332). Although coming at the issue of clustering from a different perspective (that is, of marketing rather than research), Mooi and Sarstedt come to a similar conclusion concerning validation, stating that clusters are only useful if they discriminate groups on non-clustering variables (343). Thus, for this analysis, the main forms of validation were:

- Differentiation: do groups vary significantly by variables not used for clustering? As discussed, due to the large sample size, inferential statistics tend to report even small differences as statistically significant; therefore, differences were assessed primarily on the basis of effect size (297).

- Reliability or internal consistency: was the segmentation reproducible when the file was split?

- Group size: are the groups large enough to support investment into tailoring? Having several small groups could be inefficient for organizations to invest in tailoring efforts.

- Face validity: do the groups produced conform to what is known about the HRA population? In this analysis "face validity" (231) was determined by subjectively evaluating the extent to which groups appear to conform to what is learned about the HRA population through descriptive statistics. As discussed by Harle *et al.* (229) and Dolnicar (333, 341), choosing an appropriate segmentation can be a subjective process. Harle *et al.* acknowledge that in many cases decision are made on the basis of the "subjective interpretability of the clusters" (229).

- Utility of the solution or whether the segmentation provided new information or insights about the HRA users above and beyond that obtained through analysis by demographics. This form of validation reflects Dolnicar's description of the practical

need to consult with management in choosing the sort of segmentation that meets the needs of a program or project (341).

Studying further etool engagement

Chapter 8 looks at the proportion of HRA users who enroll for one of the follow up etools (the HWAP, BPAP or eSupport). The primary objective of this chapter is to test whether the optimal segmentation developed in Chapter 7 can predict further etool engagement.

Summary

The HRA constitutes a large and diverse database, providing information on users' modifiable and non-modifiable risk factors, demographics, disease-related factors and a proxy for socioeconomic status (education). These data are not without limitations. As an open access etool, there are no experimental controls by which to verify either the validity or reliability of self-reported data. Although question phrasing and order were consistent throughout the data collection period, no research was conducted to determine what, if any, effect they could have response accuracy. Perhaps more importantly, as secondary analysis of an operational data base, corporate concerns outweighed the needs of science. As a result, the study was limited to those variables collected for the etool operation. There was no opportunity to add questions for research purposes.

The validity of self-reported data is not the only challenge in analyzing the HRA database. Most of the data points are categorical and the large size of the database limits the utility of inferential statistics in determining whether relationships or differences are meaningful. Strategies to overcome these challenges included recoding to create counts (e.g., the number of modifiable risk factors) and the use of effect size to identify which differences are substantive (i.e., medium- or large-sized effects).

Segmentation is commonly used in social marketing and the analysis of large data sets. In this analysis, three types of segmentation procedures that are readily available to non-specialists and different combinations of clustering variables will be used. The analysis will also focus on what is perhaps the greatest challenge in segmentation: determining which of several possible solutions is more robust and useful for program operators. This will require a combination of objective (quantitative) and subjective analyses.

5. Overview of the HRA Population

As shown in an earlier chapter, early in the online process there is considerable self-selection among HRA visitors. Nothing is known about visitors who do not choose to start or complete the HRA; analysis can only be conducted on the sub-set of visitors who complete it. This chapter will look at this sub-set in order to determine the characteristics of HRA completers. What sort of person completes the HRA?

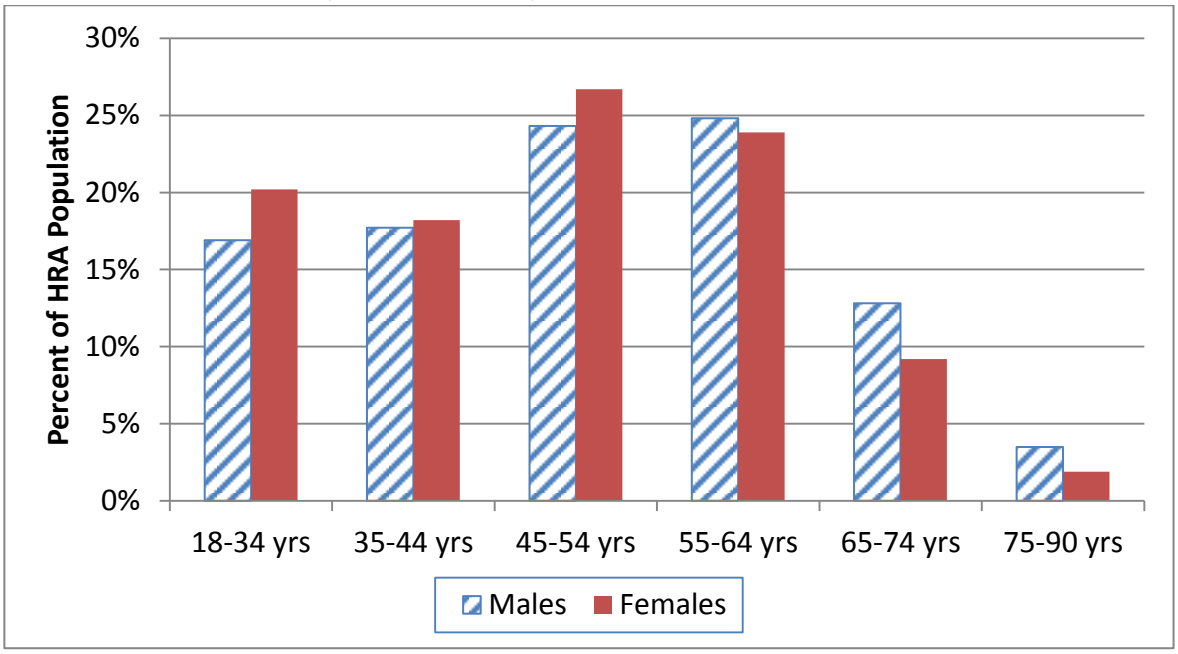
Demographic variables

Over two-thirds (68.0%) of HRA participants were female (see Table 1 in Appendix 3). The mean (sd) age of participants was 48.57 (14.14) years, with female participants being slightly younger than males (47.9 [13.9] compared to 50.0 [14.5] years). Although an independent t-test found the difference statistically significant ($p<.001$), the size of the effect of gender on mean age was small (Cohen's d=0.150). As shown by Figure 4, there were relatively few differences in the proportion of males and females by age group.

The HRA population was skewed by age, with relatively small proportions being 65 years or older (Figure 4). Kurtosis, a measure about the height or "peakedness" of a distribution (300), was strongly negative (for males, kurtosis = -0.865, standard error = 0.025 and for females kurtosis = -0.843, standard error = 0.017), suggesting a relatively flat and weak-tailed distribution. Skewness was .092 (standard error=.023) for males and .127 (standard error=.015) for females, suggesting only weak bias towards the younger age groups. Although kurtosis values suggest the distribution does not conform to the classic bell curve of the normal distribution, it does not mean that parametric statistics cannot be used; as discussed by Pallant (pg. 56), an abnormal distribution may not be a barrier when sample size exceeds 200 (302).

Participants tended to be well educated, with a total of 75.3% reporting some post-secondary education (15.3%) or having graduated from a college or university (60.0%). Only a minority of users reported not completing high school (4.4%) or no post-secondary education (18.7%). Excluding those who did not want to give their highest level of education (n=1,693), there was a large effect (ϖ=.143) for mean age to decrease with education level, being 55.1 (14.8) years for those with less than a high school education, 51.0 (15.0) for those with high school, 48.9 (14.3) for those who did not graduate from college or university and 47.3 (13.5) for those who graduated.

Figure 4: Proportions by Age Group by Gender



The majority (83.5%) of those who completed the HRA gave their ethnicity as white or Caucasian. The three most common non-Caucasian ethnic groups were "other" (4.3%), Chinese (3.3%) and South Asian (2.6%). The small size of most ethnic groups suggests they may not be efficient basis on which to tailor messaging (see Table 1 in Appendix 3). Even when the three ethnic groups thought to be at increased risk of CVD were combined (i.e., South Asians, Aboriginal Canadians, and people of Black or African descent), they accounted for only 5.5% of users.

Over 60% worked in "white collar" occupations such as management, health or education and close to 60% were employed full or part-time (57.5%) or were married (58.3%). Excluding those who did not report their type of occupation (n=18,136), mean age was 47.4 (13.3) for those working in "white collar" occupations, 45.8 (14.4) for those in sales or service, and 48.7 (13.0) for those in the trades, but the effect was small (ϖ=.05).

For most demographic variables, age group was associated with larger effect sizes than gender (Table 1 and 2 in Appendix 3). For example, age group had a large positive effect on marital status (i.e., rates of being married or having a common-law spouse rose with increasing age: *eta* = .271), a large negative relationship with full- or part-time employment (*eta*=.157), and a large but non-linear relationship with white collar employment (*eta*=.156) and post-secondary education (*eta*=.157; Table 2 in Appendix 3). Although the proportion of users who were male was highest for the older age group (46.9% for those 75-90 years), age group had only a medium-sized effect on gender (*eta*=.084).

Non-modifiable risk factors

Overall, the HRA population tended to report a high level of non-modifiable risk factors. Over half (58.0%) of HRA participants reported a family history of hypertension, 44.8% a family history of diabetes, 44.7% dyslipidemia and 38.5% premature heart disease. Less common non-modifiable risk factors were a family history of premature stroke (15.7%) and high-risk ethnicity (5.5%).

Women had a higher mean number of non-modifiable risk factors than males (2.2 [1.4] *vs.* 1.9 [1.4]) but there were only small differences in the prevalence of individual risk factors. To see the rate of individual non-modifiable risk factors by gender, please refer to Table 3 in Appendix 3. None of the effect sizes for gender on the prevalence of non-modifiable risk factors met the Cramer's V_{1df} cut-off for even a medium-sized effect

Age group had a medium-sized inverse effect on three of the six non-modifiable risk factors: high-risk ethnicity (*eta*=.102), family history of diabetes (*eta*=.074) and of dyslipidemia (*eta*=.071). Age group had only a small and non-linear effect on a family history of hypertension (*eta*=.041), premature heart disease (*eta*=.032) and stroke (*eta*=.031). (To see all rates of all variables by age group, please refer to Table 4 in Appendix 3.)

There was no difference in the mean number of non-modifiable risk factors for those with lower compared to higher education (for both groups mean=2.1, with sd of, respectively, 1.4 and 1.5; $p<.001$ but $\varpi = .01$). There was also virtually no difference by type of occupation, being 2.1 (1.4) for those in white collar occupations, 2.1 (1.5) for those in sales or service, and 2.0 (1.5) for those reporting they work in trades ($\varpi=.02$).

Modifiable risk factors

A poor diet, being overweight or obese, and physical inactivity were reported by half or more of HRA users (respectively, 69.5%, 56.1% and 49.4%). Other less common risk factors were high salt consumption (35.6%), drinking in excess of the low-risk drinking guidelines (24.1%) and frequent stress (19.7%). The least frequently-reported modifiable risk factor was smoking (12.5%).

Readiness to change varied between different risk factors. For example, 35.6% of those who were overweight or obese said they were not willing to change for at least the next six months (i.e., were in the Precontemplation stage) compared to 24.0% of smokers

and 14.1% of those who were physically inactive (for more information, please refer to Table 5 in Appendix 3).

There was no difference by gender in the mean number of modifiable risk factors: for both mean=2.6 with a standard deviation of 1.6 (Cohen's *d*=0.058, indicating a small effect). As shown in Table 5 in Appendix 3, there were some small differences between men and women in the type of modifiable risk factors reported. Females were more likely than males to report physical inactivity (52.4% *vs.* 43.1%, Cramer's V_{1df}=.091) and frequent stress (22.6% *vs.* 13.6%, Cramer's V_{1df}=.105). For all other risk factors, males had higher rates. Males had a higher rate of overweight/obesity (65.5% of males *vs.* 60.1% of females; Cramer's V_{1df}=.052), excess alcohol consumption (31.8% *vs.* 20.5%; Cramer's V_{1df}=.124), and poor dietary behaviours (72.3% *vs*. 68.3%; Cramer's V_{1df}=.041). Men were slightly less likely to say they try to reduce the amount of salt they eat (50.3% compared to 55.4% for women) and were more likely to say they don't monitor or control their salt intake (43.2% *vs.* 9.8%; Cramer's V_{1df}=.055). The only modifiable risk factor for which there was no difference by gender was smoking: 13.0% of males and 12.4% of females were smokers (Cramer's V_{1df}=.007, *p*=.013). In no cases did gender have even a medium-sized effect on the reporting of modifiable risk factors (for all, Cramer's V_{1df}<.30).

In summary, males and females had similar numbers but different types of modifiable risk factors. However, in most cases differences by gender were relatively small, making gender a problematic variable for tailoring risk factor messaging.

Table 6 in Appendix 3 shows the prevalence of modifiable risk factors by age group. As it shows, age group had a large inverse effect on the report of high salt consumption (*eta*=.214), poor dietary behaviours (*eta*=.139), and frequent stress (*eta*=.144) and a medium-sized inverse effect on physical inactivity (*eta*=.095) and smoking (*eta*=.091). There was a medium-sized effect for the prevalence of overweight/obesity (*eta*=.070) but the relationship was not strictly linear. The prevalence of modifiable risk factors appeared to consistently decrease as age increased, with the exception of excess alcohol consumption, for which there was no effect by age group (*eta*=.006).

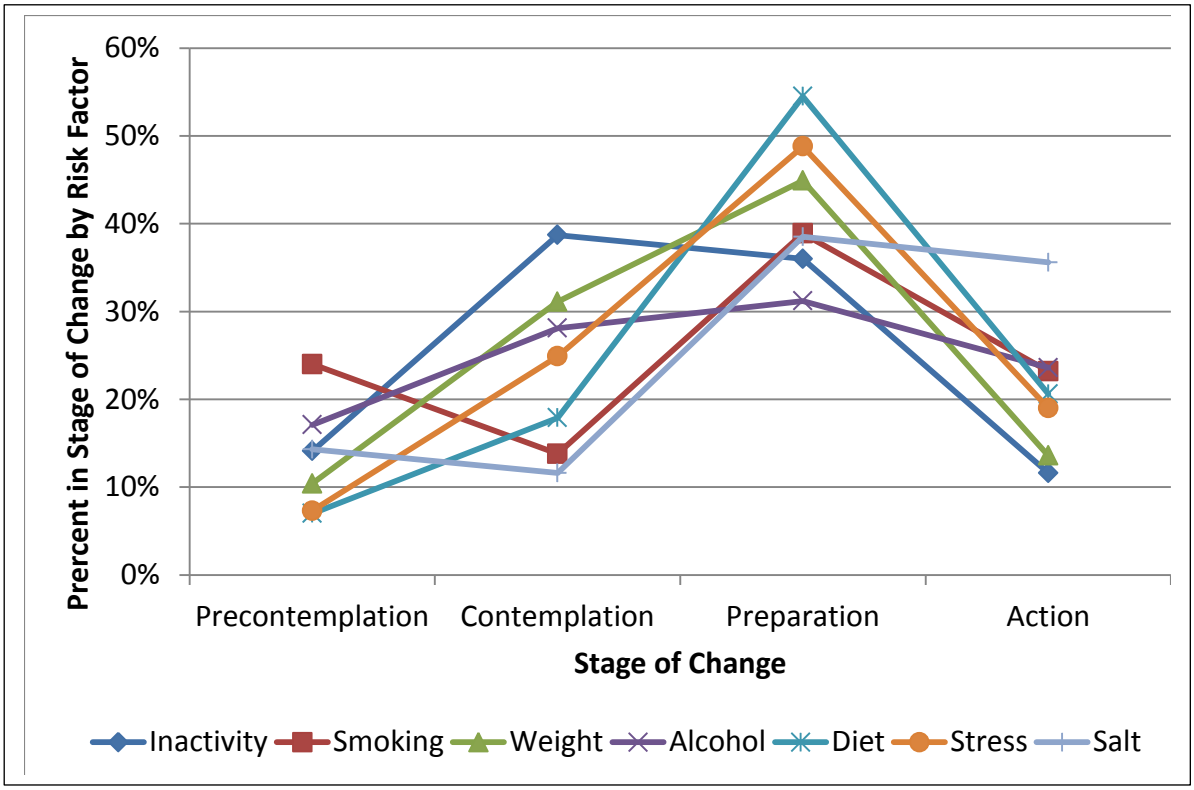
The mean number of modifiable risk factors was moderately higher for those without a post-secondary education compared to those with more education (2.7 [1.4] vs. 2.5 [1.4], ϖ=.06, a medium-sized effect). Mean number of modifiable risk factors was similar for those working in sales or service (2.8 [1.4]) and trades (2.8 [1.4]) and lower for those reporting "white collar" occupations (2.6 [1.4]), which was associated with a medium-sized effect (ϖ=.08). The mean was also higher for those belonging to a high-risk ethnic group compared to those who did not (2.8 [1.4] *vs.* 2.6 [1.4]) but the effect was

smaller (ϖ=.05). When age, gender, education level (high/low) and ethnicity (high/low risk) were entered into a univariate regression (occupation was not included as it is probably confounded by education), they explained only 4.6% of variance (adjusted r^2=0.46, p<.001). This suggests that although education and ethnicity played some role in the reporting of modifiable risk factors, the effect was small.

Readiness to change modifiable risk factors

When a modifiable risk factor was reported, users were asked when they would be willing to make changes to address it, thus documenting their readiness to change. Figure 5 shows the proportion of those with a risk factor who placed themselves in the four stages of change. As it shows, for all risk factors except physical inactivity the most commonly-reported stage of change was Preparation, with the highest being 54.5% for those reporting a poor diet and lowest for alcohol consumption at 31.2%. There were differences between risk factors in the proportions reporting themselves either willing or unwilling to change. For example, a quarter (24.0%) of smokers reported they were in the Precontemplation stage, compared to a low of 7.0% for those reporting a poor diet.

Figure 5: Proportions in Stage of Change by Risk Factor



Of the seven risk factors, the greatest willingness to change, as indicated by being in either the Preparation or Action stage, was for diet (75.1%) and salt consumption (74.0%), followed by stress management (67.8%), smoking (62.2%), weight (58.5%),

alcohol (54.8%) and physical activity (47.2%). Neither gender (see Table 5 in Appendix 3) nor age group (Table 7 in Appendix 3) had a large effect on the distribution of the stages of change.

Are the proportions reported in the HRA similar to those observed in similar populations? The answer may vary according to the type of risk factor. For example, data from an American study of primary care patients at increased risk of coronary heart disease found that at baseline approximately a quarter of respondents were in the Precontemplation stage for reducing dietary fat (345). In the HRA, however, only small proportions reported frequent consumption of fatty foods (13.0%) or fast foods (2.9%), and of those who did, only 7.0% were in the Precontemplation stage for dietary change. Likewise, in the U.S. study, about a third were in the Precontemplation stage for physical activity (345), compared to 14.1% in the HRA. Finally, in the U.S. study 39.3% of smokers were in the Precontemplation stage (345), compared to 24.0% of those in the HRA.

Other studies of stage of change for physical activity have cited proportions in the Precontemplation stage ranging from 29.6% (346) to 8% (347). The distribution of willingness to change may vary significantly by not only risk factor but the population being surveyed and how the question is asked. However, from even this cursory review of the literature it appears HRA respondents demonstrated a greater readiness to change. For all of the modifiable risk factors, less than a third of HRA respondents indicated they were not willing to consider behaviour change (i.e., placed themselves in the Contemplation or Precontemplation stages), although the proportion varied according to the individual risk factor (see Tables 5 and 7 in Appendix 3).

Figure 6: Proportion of HRA population in contemplation or precontemplation stage of change by modifiable risk factor and age group

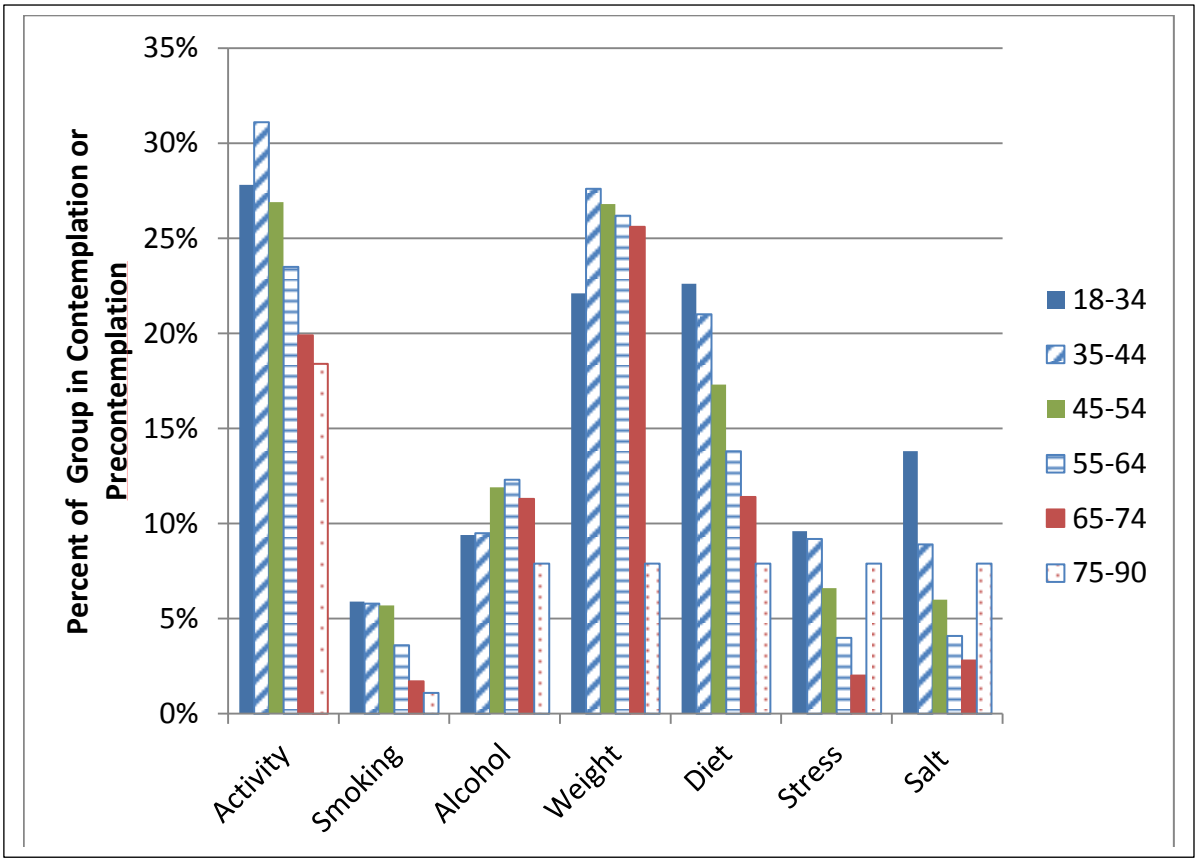
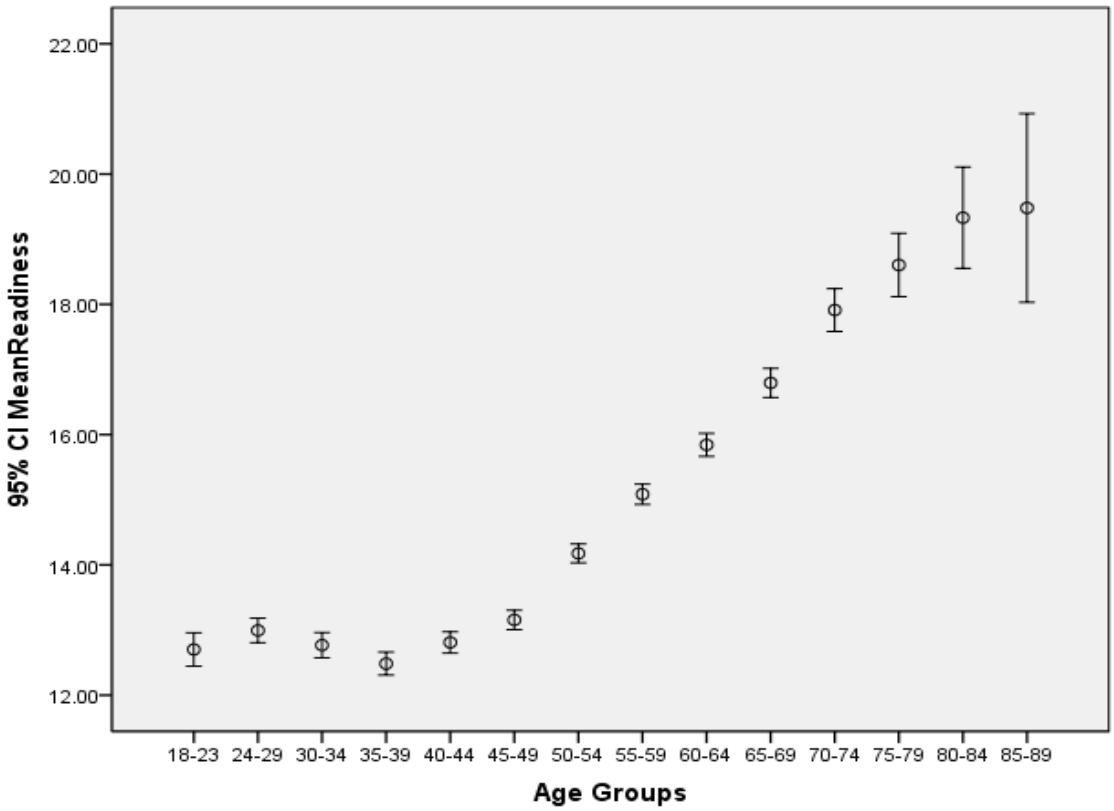


Figure 6 shows the proportion of users by age group who were not ready to change for each of the seven modifiable risk factors. There was a moderate tendency for unwillingness to change stress (*eta*=.097), salt consumption (*eta*=.093), diet (*eta*=.081) and alcohol consumption (*eta*=.074) to vary by age; for activity, smoking and weight effect sizes by age were small. Of the moderately strong relationships only diet was linear, suggesting that for most risk factors there may be complex relationships between age and readiness to change. When a lifestyle healthiness score was calculated to reflect the stage of change for all of the seven modifiable risk factors it had a J-shaped relationship with age (see Figure 7). Note that as the size of the age groups decreased, the margins of error became larger.

Figure 7: Mean Readiness to Change (Healthiness Score) by Age Group



Excluding non-respondents, the overall healthiness score tended to increase with education, the means being 28.1 (4.1) for those with less than high school, 28.7 (3.9) for high school graduates, 28.5 (3.9) for those who didn't graduate college or university, and 29.2 (3.8) for those who graduated, with the effect being medium-sized (ϖ=.09). There was a medium-sized effect (ϖ=.07) for those working in "white collar" occupations to have a higher healthiness score compared to those working sales or service or the trades (respectively, 29.0 [3.8]), 28.3 [4.1]) and 28.6 [4.0]).

Prevalence of chronic diseases

Over a quarter (26.1%) of respondents reported being told by a health professional they have hypertension or had been prescribed blood-pressure-lowering medication. The rate was higher among males than females (32.5% vs. 23.1%) but the effect of gender was small (Cramer's V_{1df}=.099; see Table 8 in Appendix 3). The next most common diagnosis was dyslipidemia, reported by 27.9% of males and 17.5% of females (Cramer's V_{1df}=.120, a small effect). The least common CVD-associated chronic conditions reported by users were diabetes (9.5% of males and 5.6% of females; Cramer's V_{1df}=.071), heart disease (7.7% of males and 2.9% of females; Cramer's V_{1df}=.109) and stroke (2.7% of

males and 1.8% of females; Cramer's V_{1df}=.030). For all of these conditions, gender had only small effects.

Half (51.9%) of the women and 65.7% of the men reported none of the six vascular-related conditions, for an overall total of 61.3% (Cramer's V_{1df}=.152, a small effect). This finding suggests half of the women and two-third of the men who come to the HRA may be more interested in prevention than chronic disease management.

The prevalence of not only vascular conditions but most chronic conditions, excepting asthma and mood disorders, increased with age (see Table 9 in Appendix 3). Effect of age upon the report of chronic conditions was large for hypertension (*eta*=.338), arthritis (.289), dyslipidemia (.282), osteoporosis (.216), heart disease (.172), and diabetes (.148), and moderate for cancer (.119), COPD (.095), sleep apnea (.094), stroke (.087), back pain (.078), mood disorder (.063), and asthma (.056).

Excluding those who did not give their highest level of education, there was a medium to large effect (ϖ=.132) for the mean number of vascular conditions to have an inverse relationship with highest level of education, being 1.04 (1.2) for those with less than a high school education, 0.74 (1.03) for those with high school, 0.67 (1.00) for those who did not graduate from college or university, and 0.53 (0.89) for those who had graduated. For occupation, there was a medium-sized effect (ϖ=.07), with the means being 0.54 (0.90) for white collar occupations, 0.58 (1.00) for those working in sales or service, and 0.76 (1.06) for those working in trades.

Among the non-CVD-related chronic diseases, the most common were arthritis (12.7% of males and 19.0% of females) and mood disorders (11.3% of males and 19.6% of females). The HRA may be particularly effective in attracting people with CVD-related conditions such as hypertension (26.1%) and dyslipidemia (20.8%), rather than other common conditions that are not necessarily associated by the general public with CVD, such as arthritis (18.0%), mood disorder (16.9%), sleep apnea (5.7%) or COPD (5.6%).

Given the increase observed in the prevalence of most chronic diseases with age, it was not surprising that the report of taking any form of prescription medication showed a similar trend (see Table 9 in Appendix 3). Report of being prescribed medication increased from 22.0% for those 18-34 years to 74.2% for those 75-90 (*eta*=.301, a large effect). In logistic regression, age in years and number of vascular diseases explained between 22% and 30% of variance (Cox and Snell r^2=.224 and Negelkerke r^2=.305) and increased the proportion of cases correctly categorized from 61.3% (beginning block) to 76.9% (model X^2 <.001). Each additional vascular disease reported increased the odds or being prescribed medication three-fold (OR=3.220, 95% CI 3.158-3.285; β[SE] =1.170

[.010], Wald = 13514.48 with 1df, p<.001), while the odds increased 2% for each year of age (OR=1.024, 95% CI 1.023-1.035; β[se]=0.24 [.000), Wald=3107.61 with 1 df, p<.001). Although these effects were statistically significant (model X^2 <.001), the Hosmer and Lemeshow X^2 was <.001, suggesting less than ideal fit (301). Adding gender did not improve the model, as indicated by the percentage of cases correctly predicted, amount of variance explained or Hosmer and Lemeshow goodness of fit.

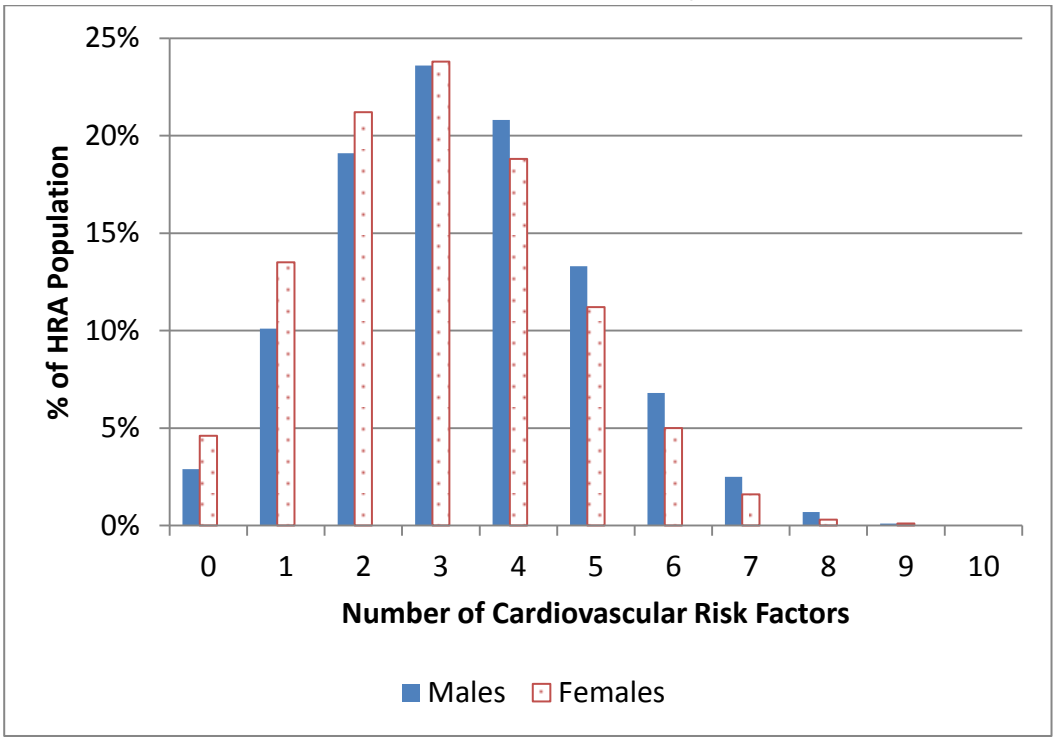
Among those prescribed medication, the proportion reporting they miss it some or most of the time decreased with age, from 22.1% for those aged 18-34 years to 5.6% among those 75-90 (*eta*=.143, a large effect; see Table 9 in Appendix 3). In logistic regression, age in years had a 3% negative effect on poor medication adherence (OR=.969, 95% CI .968-971, β[SE]=-0.931 [.001], Wald = 1592.78 with 1df, p<.001; model X^2 <.001) but the amount of variance explained was small (Cox and Snell r^2=.029 and Nagelkerke r^2=.053) and the model had poor fit (Hosmer and Lemeshow<.001). Adding gender or number of vascular diseases did not change the size of the effect by age, the Hosmer and Lemeshow X^2, or increase either the proportion of cases correctly predicted, or the amount of variance explained (r^2).

Total number of cardiovascular risk factors

Total number of CVD risk factors was calculated based on the sum of modifiable risk factors (physical inactivity, overweight/obesity, unhealthy diet, excessive salt consumption, high-risk alcohol consumption, smoking, frequent stress) and the report of three medical risk factors (hypertension, diabetes, dyslipidemia). The mean number of CVD risk factors was higher for males (3.34 [1.65]) than females (3.03 [1.62]); although this differences was statistically significant (p<.001 when tested using an independent t-test) the effect size was small (Cohen's d=.190). Only a small proportion (2.9% of males and 4.6% of females) reported no CVD risk factors (see Table 10 in Appendix 3).

Figure 8 shows the number of risk factors reported by gender. There were relatively small differences between males and females; for both there were uni-modal bell curves skewed to the left (i.e., towards a lower number of CVD risk factors). Kurtosis values are negative but close to zero (for males, -0.148, SE=0.025 and for females -0.181, SE=0.017).

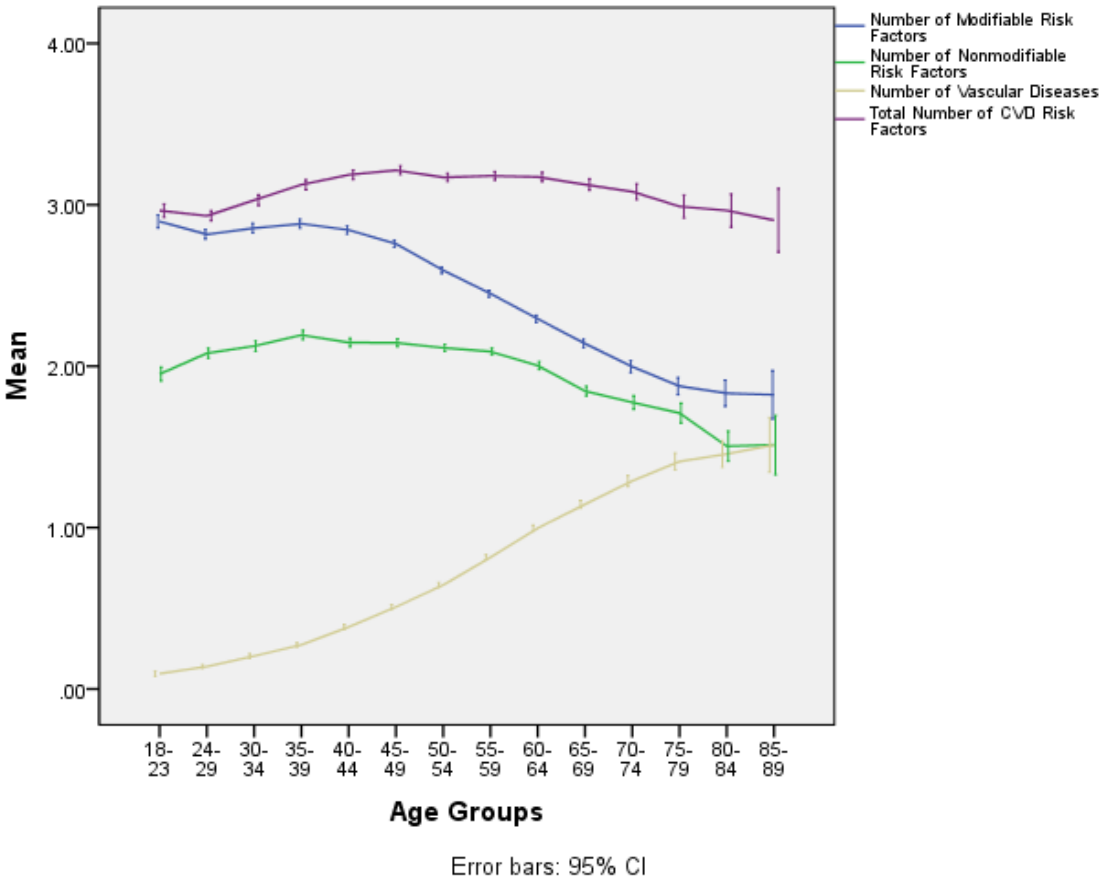
Figure 8: Number of cardiovascular risk factors by gender



There was evidence that the type of CVD risk factor reported by users varied by age. Table 11 in Appendix 3 shows the mean number of modifiable and non-modifiable risk factors, vascular conditions and total CVD risk factors (sum of modifiable risk factors and vascular conditions of hypertension, diabetes and dyslipidemia) by age group. There was a significant and large effect of age on mean number of vascular conditions (ϖ =.369) and number of modifiable risk factors (ϖ =.200). However, age group had only a moderate effect on number of non-modifiable risk factors (ϖ =.079).

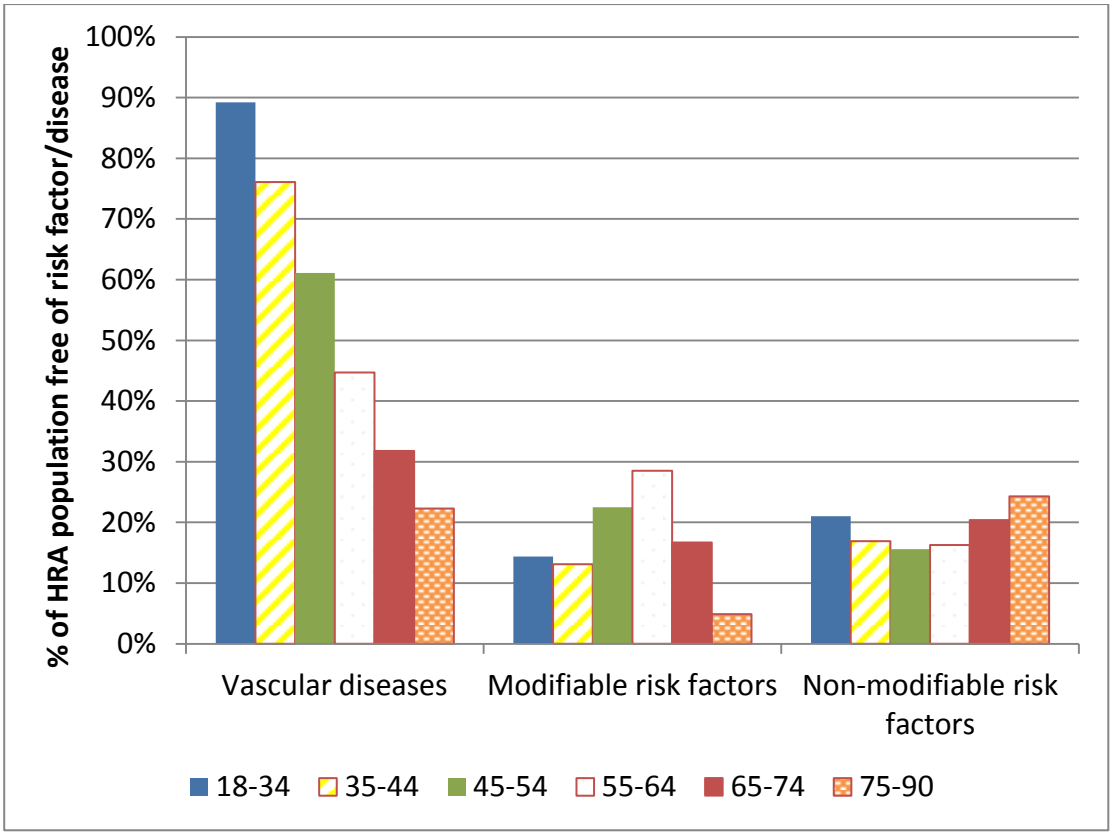
Means by five-year age groups are illustrated in Figure 9. As it shows, although the number of vascular diseases increased with age the number of modifiable risk factors decreased substantively and the number of non-modifiable risk factors more modestly. As a result, the total number of CVD risk factors remained fairly consistent across the age groups.

Figure 9: Mean number of total CVD risk factors, modifiable risk factors, non-modifiable risk factors and vascular diseases by age group



Do the trends shown in Figure 9 suggest that age may be a viable method for grouping and tailoring HRA messaging? To explore this possible, the proportion of HRA users who were free of the three types of CVD risk factors were analyzed by age group (Figure 10). It suggests there are limitations to tailoring messaging based on age. For example, although older users were more likely to report vascular diseases, substantive proportions (roughly 30% of those aged 65-74 and 20% of those 75-90) actually had no conditions. In addition, 10% of the youngest age group (18-34) and a quarter of those 35-44 had vascular conditions. In other words, assuming that all older participants have vascular conditions while no younger participants are affected would misclassify a substantive number of users. Likewise, there were no clear age-related trends in the reporting of modifiable and non-modifiable risk factors. In summary, the proportions shown in Figure 10 suggest that age may be an imprecise method of tailoring program messages.

Figure 10: Proportion of HRA users who are free of vascular health concern by type and age group



Vascular Disease Management

Hypertension, dyslipidemia and diabetes are important risk factors for CVD and are captured by the HRA. For those without a condition, a follow-up question addressed the frequency of screening. Those who report a condition were asked follow-up questions about medications, testing and control. These questions may afford insights into the health behaviours of HRA users.

Hypertension management

Detailed information on hypertension management by gender is provided in Table 12 and by age group in Table 13 in Appendix 3.

Almost three-quarters (73.9%) of HRA users said they had not been diagnosed with hypertension or take blood pressure-lowering medication. The majority (79.2%) of these users had their blood pressure taken by a health professional (i.e., were screened for hypertension) within the past 12 months. Only small proportions reported they had not

been screened within the past two years (5.7%), had never been screened (1.5%) or did not know when they were last screened (1.7%). Gender had only a small effect on how frequently normotensives were screened (Cramer's V_{1df}=.084) but age group had a large effect (*eta* =.211).

A quarter (26.1%) of HRA users reported a diagnosis of hypertension. Of these, three-quarters (75.1%) had been prescribed medication for their condition. There was a medium-sized effect for more women than men to be prescribed medication (61.0% vs. 39.8%, Cramer's V_{1df} =.029). Age group had an even larger effect, with the rate increasing from 39.0% for those aged 18-34 to 84.7% for those 75-90 (*eta*=.214). Over three-quarters (77.7%) of those with hypertension reported their blood pressure had been checked by a health professional within the past 6 months. There was no strong difference in blood pressure testing by gender (females = 78.5% *vs.* 76.5% for males; Cramer's V_{1df} =.032) but there was a moderate effect by age group (*eta*=.127).

Despite the relatively high level of monitoring and the prescription of medications, blood pressure control was sub-optimal. Only half (54.2%) of hypertensives reported their blood pressure was in a healthy range most of the time. Gender had a small effect on blood pressure control (55.7% of females and 52.1% of males reported good control, Cramer's V_{1df} =.049). Good blood pressure control increased with age, from 34.9% among those 18-34 years to 74.7% among those 75-90 (*eta* =.246, a large effect).

There was no linear relationship between good control and level of education. Rates of control were 60.5% for those with less than a high school education, 63.3% for those with high school only, 59.5% for those with some post-secondary education, and 60.7% for those who had completed college or university. The effect of education was small (*eta*=.022) and non-linear (linear-by-linear association chi square *p* = .712).

Dyslipidemia management

Detailed information on dyslipidemia management by gender is provided in Table 14 and by age group in Table 15 in Appendix 3.

Almost 80% (79.2%) of HRA users had no diagnosis of dyslipidemia. Of those without dyslipidemia, about half (54.9%) said their lipids had been tested within the past 12 months. There was no significant difference by gender in the proportion screened within the past 12 months (55.2% of males and 54.8% of females; Cramer's V_{1df}=.034). Age had a strong effect on report of screening within the last 12 months, with the rate increasing from 29.5% among those aged 18-34 years to 80.8% among those 75-90 (*eta*=.412).

Approximately 20% of HRA users (27.9% of males and 17.5% of females, Cramer's V_{1df}=.120, a small effect) reported dyslipidemia. Age group had a large effect (*eta*=.282), with the prevalence increasing from 18.3% among those 18-34 years to 81.5% among those 75-90 years. Of those who reported a diagnosis of dyslipidemia, 62.2% had been prescribed medication. More men than women were prescribed medication (68.1% *vs.* 57.7%, Cramer's V_{1df}=.106, a small effect). The report of prescription medication increased strongly with age (from 18.3% for those 18-34 to 81.5% for those 75-90, *eta*=.309).

Two-thirds of those with dyslipidemia reported a lipid test within the past six months. There was only a small difference between men and women (68.2% vs. 64.2%, Cramer's V_{1df}=.042). Recent testing increased steadily from the age group 18-34 years (43.0%) to 54-74 (72.1%) but then declined modestly for the 75-90 group (69.5%). Nevertheless, age was associated with a large effect size (*eta*=.211).

The overall proportion of HRA users who reported their lipids were in a healthy range most of the time was 43.6%, which was lower than the rate of control for hypertension (54.2%). Men were somewhat more likely to report good lipid control than women (47.0% vs. 41.0%, Cramer's V_{1df}=.087). Age had a stronger effect on control rates: the proportion reporting good control increased from 25.2% among those 18-34 years to 68.7% for those 75-90 (*eta*=.265).

Rates for good control were 52.2% for those with less than a high school education, 50.3% for those with high school only, 47.8% for those with some post-secondary education and 48.3% for those who had complete college or university. The effect of education on good lipid control was small (*eta*=.040) and non-linear (linear-by-linear association chi square *p*=.845).

Diabetes management

Detailed information on diabetes management by gender is provided in Table 16 and by age group in Table 17 in Appendix 3.

The majority (93.2%) of HRA respondents had not been diagnosed with diabetes. Among non-diabetics, 58.5% reported their blood glucose had been screened within the past 12 months. There was no difference by gender in recent screening (58.3% of males and 58.6% of females, Cramer's V_{1df}=.050). As with hypertension and dyslipidemia, screening increased with age group, from 35.8% among those 18-34 years to 79.7% for those 75-90 (*eta*=.333, a large effect).

In total, 6.8% of respondents reported diabetes: 9.5% of males and 5.6% of females (Cramer's V_{1df} =.071, a small effect). The rate of diabetes increased with age group, ranging from 1.9% for those 18-34 years to 14.5% among those 75-90 (*eta*=.148, a large effect).

Over two-thirds (69.4%) of those with diabetes reported being prescribed medication. There was a small difference in the rate of medication use by gender (71.8% of males vs. 67.5% of females, Cramer's V_{1df}=.047). The rate of being prescribed medication increased from 54.3% for those 18-34 years to 71.8% for those 55-64 but thereafter varied only minimally (*eta*=.074, a moderate effect).

A little less than two-thirds (62.8%) of those with diabetes reported a hemoglobin A1c test within the past six months. There was no significant difference by gender (62.9% *vs.* 62.8%, Cramer's V_{1df}=.032, *p*=.108). An A1c test within the past six months was reported by half (51.4%) of those aged 18-34 years and increased up until the 65-74 age group (76.2%), after which it did not increase significantly (for those 75-90 years it was 75.1%). Overall, there was a large effect of age group on the report of recent A1c testing (*eta*=.185).

Half (54.7%) of those with diabetes said their blood glucose was in a healthy range most of the time. Good glucose control did not vary significantly by gender (55.6% of males *vs.* 54.0% of females, Cramer's V_{1df}=.043, *p*=.010) but increased with age. Of those 18-34 years, 47.0% said their glucose was in a healthy range most of the time, increasing in a linear fashion to 78.1% among those 75-90 (*eta*=.186).

Good control did not vary by education. Over half (58.1%) of those with less than a high school education reported good control, compared to 62.3% of those with a high school education, 58.2% of those with some post-secondary education, and 62.6% of those who completed college or university. The effect of education on blood glucose control was small (*eta*=.027) and non-linear if a cut-off of *p*<.001 is used (linear-by-linear association chi square *p*=.030).

Summary

The objective of this chapter was to describe the characteristics of HRA users so as to add to the knowledge about open-access etool users. Analysis showed the HRA population was predominately female (68.0%), middle-aged (50.0% were between 45 and 64 years of age), well educated (60.0% graduated from college or university) and Caucasian (83.5%). These characteristics suggest that the HRA population may be

similar to the approximately 60% (51) of the general population that has been described as online health information seekers (48, 348).

Although the HRA population appeared at first glance to be relatively healthy, up to half reported one or more vascular diseases and the vast majority (96%) one or more modifiable CVD risk factor. Although the prevalence of vascular disease followed age-related trends observed in population-based surveys (349-351), trends for modifiable risk factors did not (349). This suggests that common demographics, such as gender and age, may not be helpful in accurately customizing health information.

.

6. Comparisons With Other Populations

In this chapter, three issues are addressed:

1) To what extent do Canadian HRA users reflect the general population of Canada? Are differences small and insignificant or substantive and systemic?
2) Does the use of an incentive influence who completes the HRA?
3) Are users of the open-access HRA are similar to, or different from, samples recruited for etool RCTs?

Is the HRA sample representative of the Canadian population?

According to Statistics Canada, males comprise 49.3% of the Canadian population aged 20 to 89 years. In contrast, males comprised only 32.1% of the HRA sample. In other words, compared to the CCHS population, HRA were twice as likely (OR=2.06, 95% CI 2.04-2.09) to be female. This difference meets Ferguson's requirement for a RMPE (313).

For detailed information comparing the general and HRA populations by gender, please refer to Table 1 in Appendix 4 or refer to Figure 11.

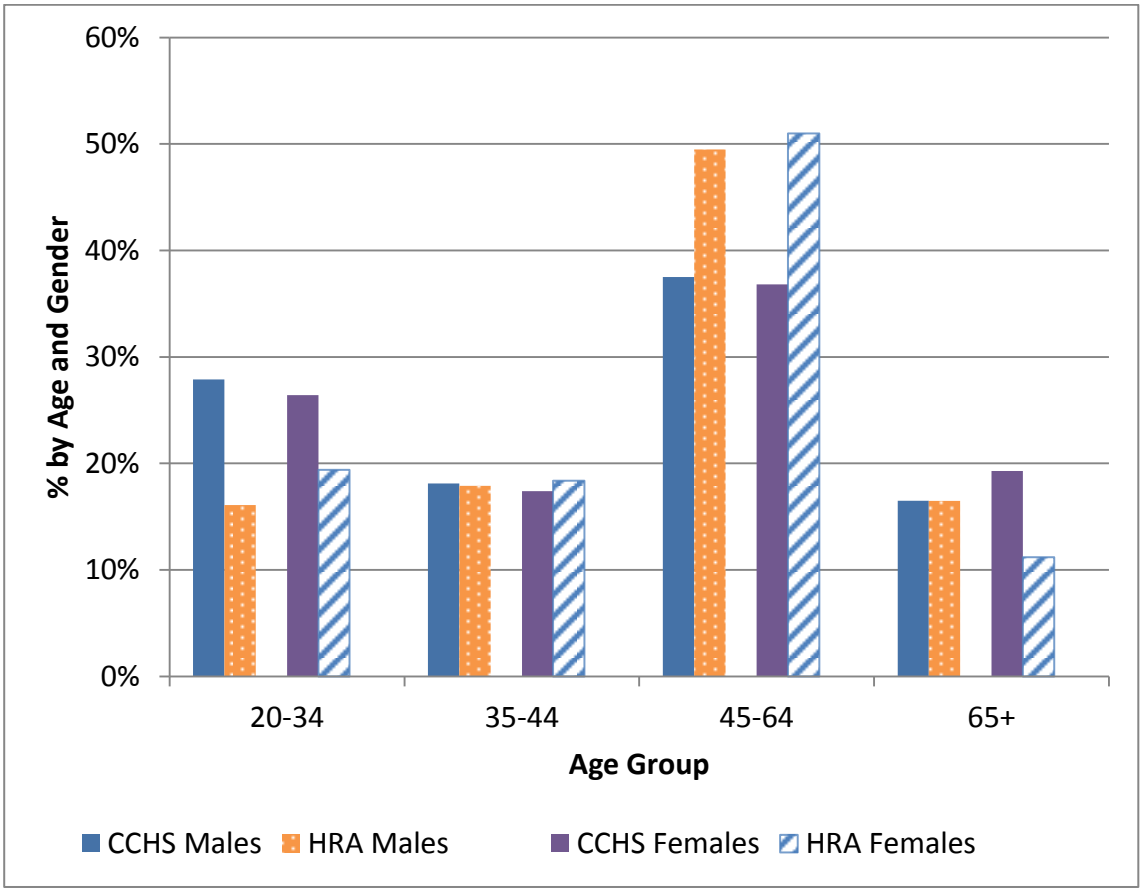
When both sexes were combined, the HRA population was 40% less likely to include adults in the 20-34 age group (OR =0.60, 95% CI 0.59-0.61, which corresponds to 1.67-fold difference). The difference was greater for young (age 20-34) males (OR=0.50, 95% CI 0.48-0.51, a 2-fold difference) than young females (OR=0.68, 95% CI 0.66-0.68 or a 1.47-fold difference).

There was little difference between the CCHS and the HRA in the proportion of participants age 35-44 (for both sexes, OR=1.03, 95% CI 1.02-1.05). As well, the HRA modestly over-represented adults 45-64, although neither the effect for males (OR=1.64, 95% CI 1.60-1.67) nor females (OR=1.78, 95% CI 1.76-1.81) met the RMPE cut-off of 2.0. When both sexes were combined, adults aged 45-64 were 73% more likely to represented in the HRA sample (OR=1.73, 95% CI 1.71-1.75).

For the oldest age group (65-89 years), the proportion of females did not vary from the CCHS sample (OR=0.99, 95% CI 0.97-1.02, p=.02). However, there were 47% fewer males in the HRA sample compared to the CCHS (OR=0.53, 95% CI 0.51-0.54). This corresponds to a 1.89-fold difference, which approaches but does not meet the RMPE cut-off.

In summary the HRA population was skewed towards women and adults aged 45-64 years of age. Compared to the general population, it appeared to under-represent younger (aged 20-34 years) males and females and females aged 65-89 years.

Figure 11: CCHS and HRA populations by age and gender



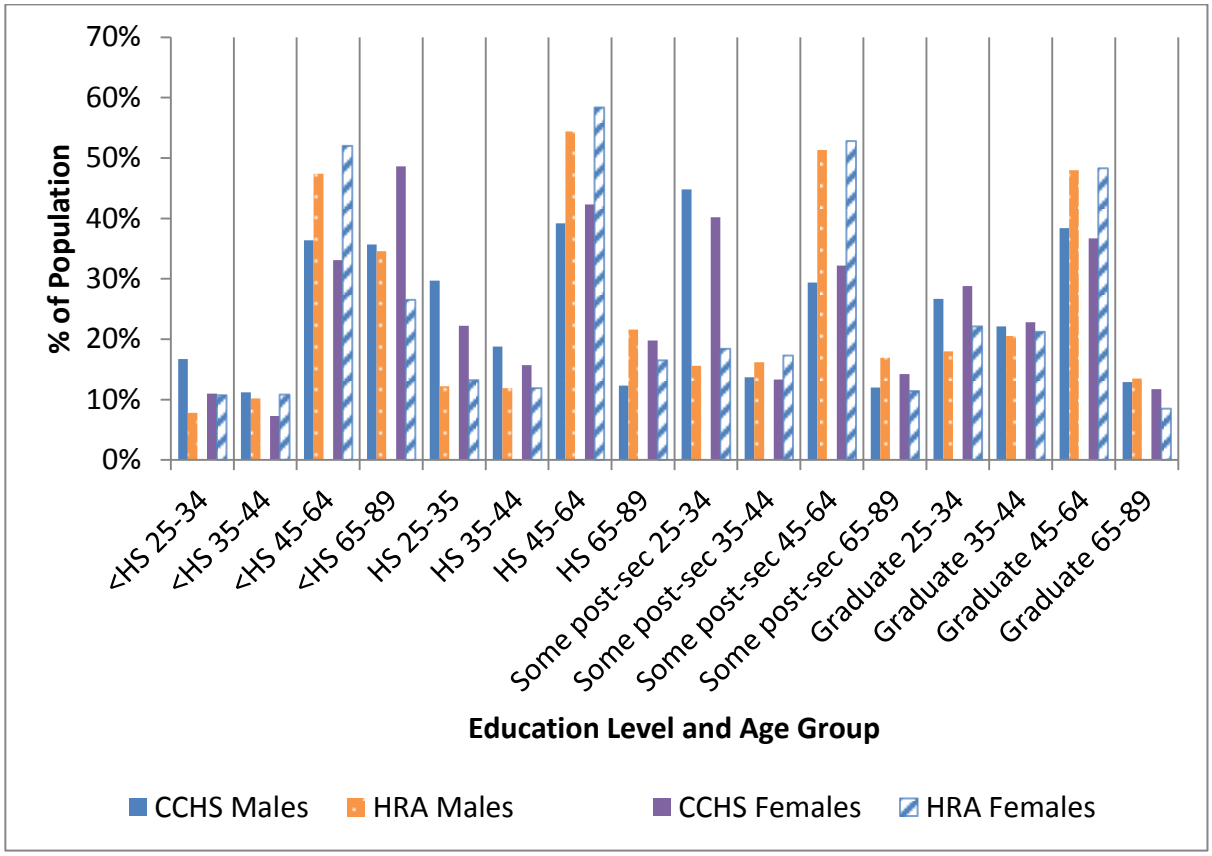
Sociodemographic variables: education

Figure 12 shows the distribution by age group and highest level of education for the CCHS population (using weighted estimates) and the HRA population (for more information see Table 2 in Appendix 4).

Compared to the general population, there was a medium-sized effect (OR=.027, 95% CI 0.26-0.38, which represents a 3.70-fold difference) for those with less than a high school education to be under-represented in the HRA sample. Effects varied by gender, with the HRA sample under-representing this level of education eight-fold in men (OR=0.12, 95% CI 0.11-0.12) in men and four-fold in women (OR=0.24, 95% CI 0.23-0.25). As shown in Table 2 in Appendix 4, the likelihood of having less than a high school education varied within gender by age group. For example, there was a 2.38-fold likelihood that males with less than a high school education would be under-represented

in the HRA (OR=0.42, 95% CI 0.36-=0.49) but a 57% increased likelihood for those 45-64 to be over-represented (OR=1.55, 95% CI 1.44-1.72). Among females, having less than a high school education was over-represented among females 35-44 years (OR=1.53, 95% CI 1.36-1.71) and 45-64 years (OR=2.19, 95% CI 2.04-2.35), but there was a 25-fold likelihood of under-representation among those 65-89 years (OR=0.04, 95% CI 0.35-0.41).

Figure 12: Comparison of CCHS and unweighted HRA populations by education, age and gender



For a high school education, overall there was no significant difference between the two samples (for both sexes, OR=1.17, 95% CI 1.15-1.19). However, in sub-group analysis the HRA significantly under-represented young males (20-35 years) with only a high school education (OR=0.33, 95% CI 0.31-0.35, representing a 3.03-fold difference), as well as those 35-44 (OR=0.60, which corresponds to a 1.67-fold difference, 95% CI 0.55-0.64). At the same time, the HRA over-represented males 45-64 (OR=1.85, 95% CI 1.76-1.95) and those 65-89 (OR=1.96, or approaching the RMPE, 95% CI 1.84-2.08).

Similarly, younger women with only a high school education were under-represented in the HRA (for 25-34 years, OR=0.53, corresponding to a 1.87-fold difference, 95% CI 0.51-0.56 and for women 35-44 OR=0.72 or 28% lower, 95% CI 0.69-0.76). At the same time, women aged 45-64 with only a high school education were over-represented, with the OR approaching the RMPE of 2.0 (OR=1.92, 95% CI 1.86-1.99).

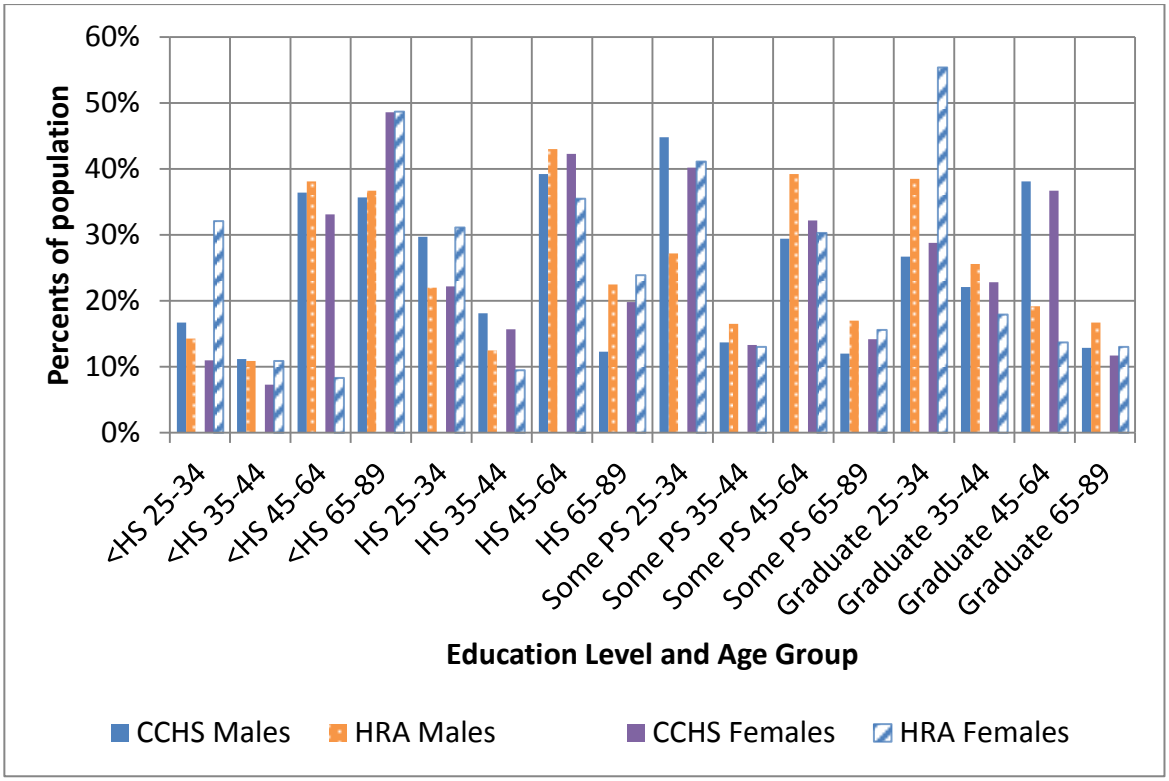
HRA users were twice as likely as CCHS respondent to report some post-secondary education (OR=2.31, 95% CI 2.27-2.35). Some post-secondary education was over-represented for males aged 45-64 (OR=2.53, 95% CI 2.41-2.67) and 65-89 (OR=1.48, 95% CI 1.39-1.59) but males 25-34 were significantly under-represented (OR=0.23, corresponding to a 4.35-fold difference, 95% CI 0.22-0.24). Likewise, middle-aged women with some post-secondary education were over-represented (OR=2.36, 95% CI 2.27-2.44) but young women 25-34 years were under-represented (OR=0.34, corresponding to a 2.94-fold difference, 95% CI 0.32-0.35).

There were no differences for either males or females in the overall proportions in the CCHS and the HRA who reported being a university or college graduate (for males, OR=0.96, 95% CI 0.94-0.98 and for females OR=1.01, 95% CI 0.99-1.03, p=0.059; for both sexes combined OR=0.88, 95% CI 0.98-1.00, p >.001). At the same time, younger (25-34 years) females in the HRA were 14-fold less likely to report being a graduate (OR=0.07, 95% CI 0.07-0.07).

Could the difference in education be due to the effect of age or gender? In a random or probability sample, all members of a population should have an equal chance of being included or selected (262). However, if certain sub-groups are over-represented, chances are unequal. Weighting attempts to compensate for this unequal sampling by adjusting the results to more closely reflect distributions in the general population (262).

Using the information from Table 1 in Appendix 4 on the distribution of the Canadian population by age and gender, post-stratification weights for the HRA sample were constructed by calculating proportion of the total population divided by proportion of the sample population. These weights were then applied to the numbers reported by education level by age group (Table 2 in Appendix 4), thereby showing the expected number of respondents in each category if there had been proportional representation by age and gender (259) (Table 3 in Appendix 4). Odds ratio were then calculated using the weighted numbers, even though testing on weighted number could compound any weighting errors (352). Please refer to Table 3 in Appendix 4 or Figure 13 for proportions when the HRA sample was weighted by age group and gender.

Figure 13: Comparison of CCHS and weighted HRA populations by education, age and gender



Even when the HRA population was weighted, it continued to under-represent those with less than a high school population. For all ages and both sexes, the OR changed from 0.27 (95% CI 0.26-0.28) to 0.26 (95% CI 0.25-0.26). This represented a 3.7% relative change, which is less than the ten percent cut-off suggested by Hernan *et al.* as an indicator of significant confounding (314). The OR of those with a high school education to be over-represented in the HRA remained small for both sexes (OR=1.08, 95% CI 1.06-1.09) and varied by only 7.7% from that reported when the HRA was not weighted (OR=1.17, 95% CI 1.15-1.19). The HRA continued to over-represent those with some post-secondary education (for both sexes and all ages OR=2.26, 95% CI 2.23-2.29), and there was only a 2.2% change from the unweighted OR (OR=2.31, 95% CI 2.27-2.35).

Weighting had the largest impact on the representation of college/university graduates. When the HRA was unweighted, there was no significant difference between the CCHS and the HRA (OR=0.88, 95% CI 0.98-1.00, p>.001). When the HRA data were weighted by age and gender, there was a small over-representation of graduates (OR=1.06, 95% CI 1.05-1.07), a 20.5% relative change. However, weighting did not change the patterns associated with the reporting of this level of education. For both the unweighted and weighted comparison, the HRA appeared to be representative of women with a college or university education (unweighted OR=1.01, 95% CI 0.99-1.03, p>.001, and weighted OR=1.10, 95% CI 1.08-1.11) while for males the HRA significant over-

represented this level of education (unweighted OR=9.57, 95% CI 9.37-9.78 and weighted OR=10.1, 95% CI 9.9-10.3). Thus, even though the relative different between ORs exceeded the ten percent cut-off, it can be argued that the small size of the ORs and the persistence of gender-specific trends suggest it is not a meaningful difference.

In summary, weighting by age and gender appeared to have little effect on the differences between the nationally-representative CCHS and the HRA in levels of education. This suggests there were real differences in the education level of the HRA sample compared to the general population of Canada. Furthermore, if education is considered a proxy for socioeconomic status, it suggests that the HRA sample is significantly different from the general Canadian population.

Comparison of health conditions or risk factors

As described, both the HRA and CCHS ask respondents about long-term or chronic conditions, with the CCHS including a qualifier of lasting, or being expected to last, for at least six months. In this section, a variety of CVD-related (diabetes, hypertension, smoking, being overweight or obese) and non-CVD conditions (arthritis, asthma, mood disorder, and chronic obstructive pulmonary disease [COPD]) are compared between the two data sources. Point prevalence from the two data sources and ORs for all eight conditions by gender and age group are shown in Table 4 in Appendix 4.

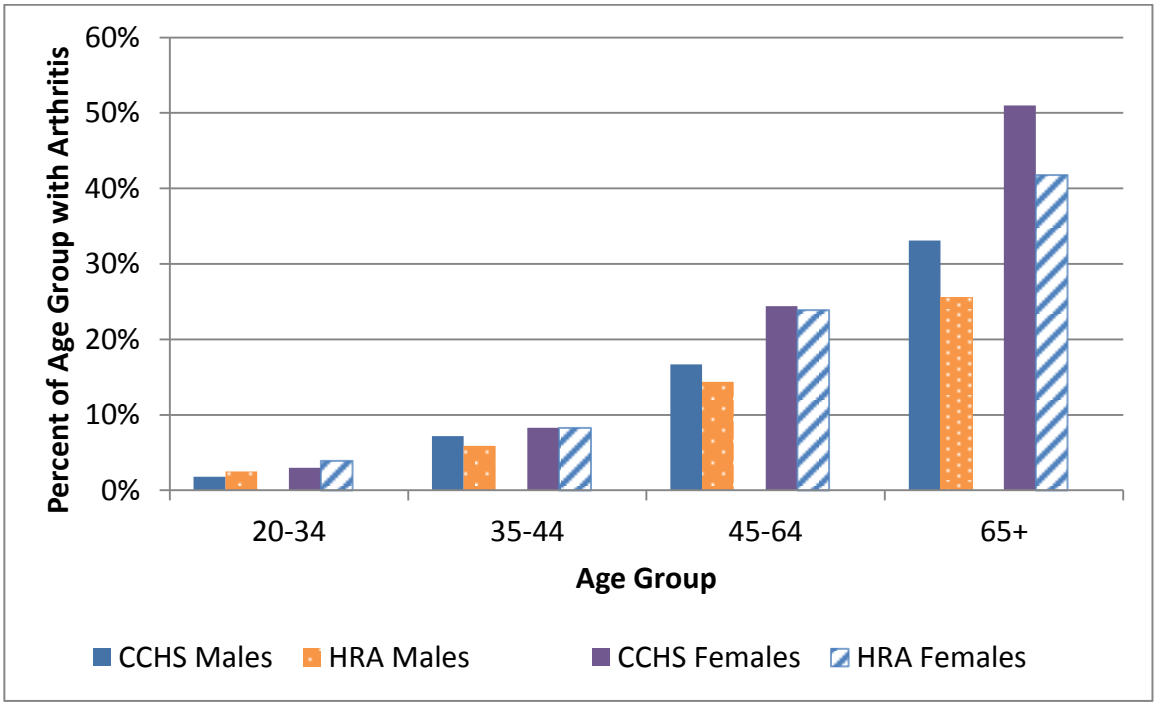
Arthritis

Arthritis, unlike chronic conditions such as hypertension or diabetes, is not a major risk factor for CVD. However, arthritis does share with CVD an age-related gradient, in that prevalence tends to increase by age (353, 354).

For all ages and both sexes combined, the HRA appeared to be representative of the population prevalence of arthritis (OR=1.02, 95% CI 1.01-1.04, *p*=.004). There were differences by age group, however (Figure 14). In the youngest age group (20-34 years), arthritis was modestly over-represented in the HRA sample for both males (OR=1.58, 95% CI 1.34-1.84) and females (OR=1.39, 95% CI 1.28-1.51). For the 35-44 and the 45-64 age groups, the HRA modestly under-represented males with arthritis (respectively, OR=0.80, 95% CI 0.72-0.88 and OR=0.87, 95% CI 0.84-0.91) but there were no significant differences for women (respectively, OR=0.99, 95% CI 0.93-1.04 and

OR=1.01, 95% CI 0.93-1.04). For the oldest age group (age 65 and over), the HRA modestly under-represented males (OR=0.75, 95% CI 0.71-0.79) and females (OR=0.75, 95% CI 0.72-0.79) with arthritis.

Figure 14: Comparison of arthritis in the CCHS and HRA by age group and gender

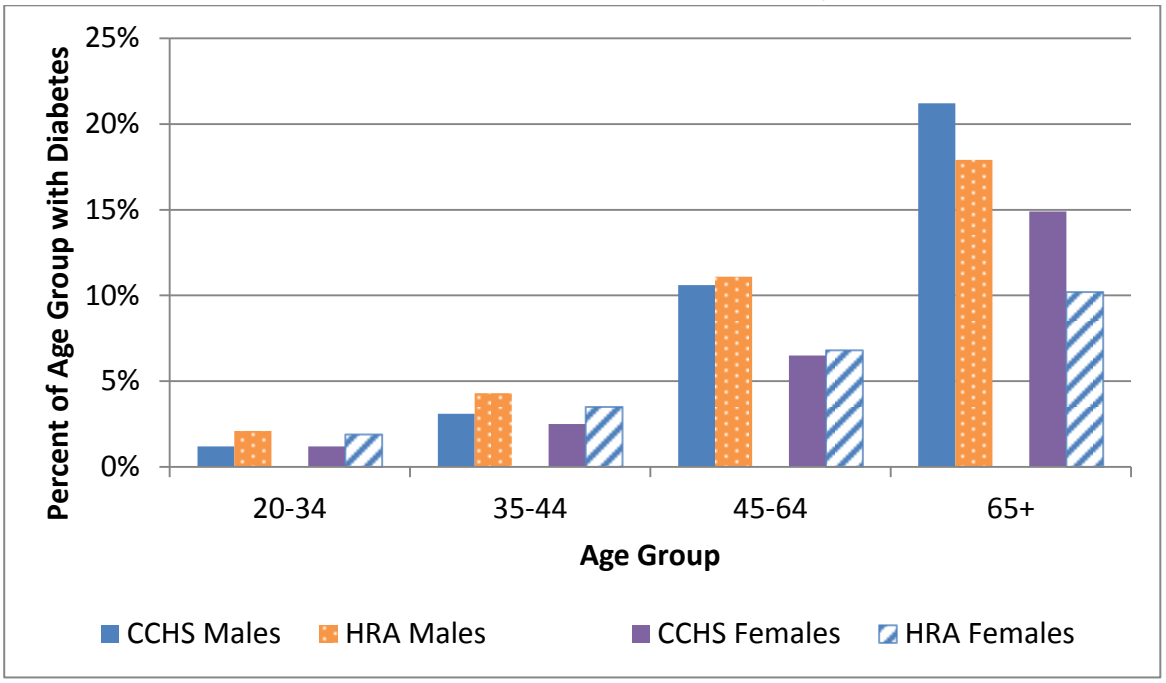


Diabetes

As a risk factor for CVD, it would be anticipated that rates of diabetes in the HRA would be higher than the general population (355). Figure 15 shows rates in the CCHS and the HRA populations by gender and age group. As it shows, there were some significant differences.

Like arthritis, diabetes was over-represented in the HRA for both men and women in the two youngest age groups, although the effect sizes were small. For the 20-34 age group, for men the OR was 1.93, approaching the RMPE (95% CI 1.62-2.30) and for women it was 1.74 (95% CI 1.44-1.95). For the 35-44 age group, the ORs were 1.36 (95% CI 1.21-1.51) for men and 1.42 (95% CI 1.30-1.55) for women. For the middle-aged group (i.e., 45-64 years), there was no substantive difference in the prevalence of diabetes for either males or females (for males OR=1.08, 95% CI 1.03-1.13 and for females OR=1.09, 95% CI 1.05-1.13). Among those 65 and over, the HRA modestly under-represented the number of people with diabetes for males (OR=0.86, 95% CI 0.81-0.92), with the effect being larger but still small for women (OR=0.68, 95% CI 0.64-0.73).

Figure 15: Comparison of diabetes in the CCHS and HRA by age group and gender

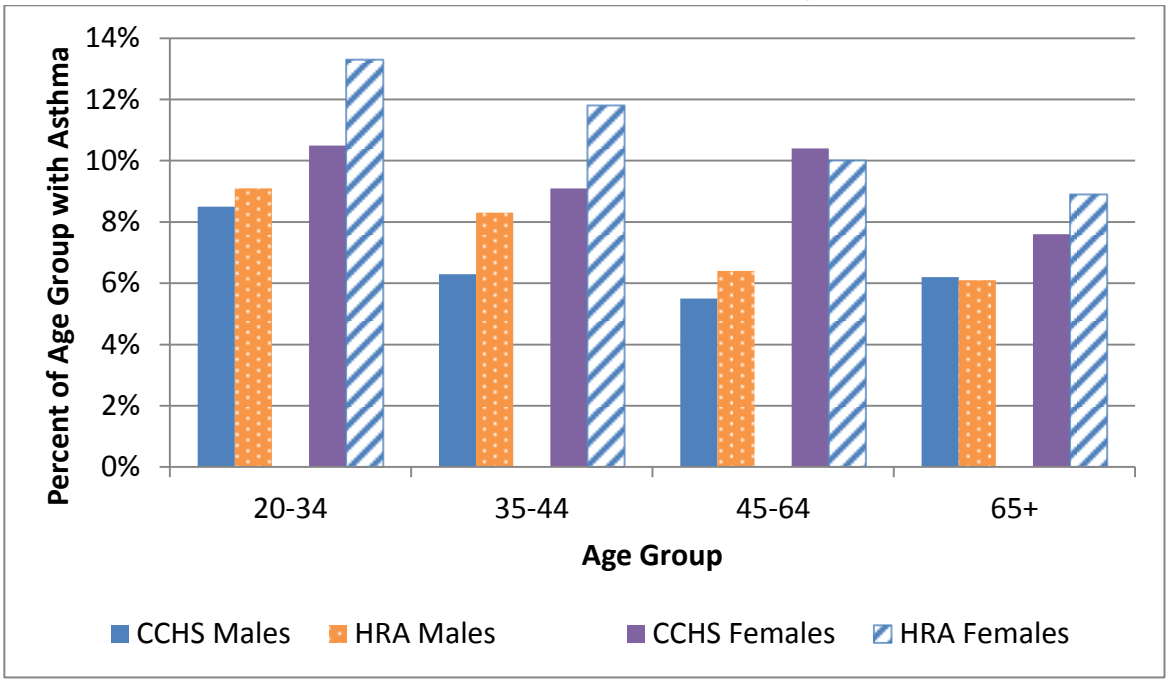


Asthma

Unlike arthritis or diabetes, the prevalence of asthma is not associated with a linear age gradient (356). As well, asthma has no apparent relationship with CVD.

Figure 16 compares prevalence of self-reported asthma between the CCHS and HRA. As it shows, for the 20-34 age group there was a small effect for asthma to be over-represented in the HRA, more so for females (OR=1.40, 95% CI 1.34-1.46) than males (OR=1.17, 95% CI 1.08-1.28). For the 35-44 age group, the HRA modestly over-represented the number of people with asthma, with the effect being similar for males (OR=1.33, 95% CI 1.22-1.45) and females (OR=1.31, 95% CI 1.25-1.38). For the middle-aged group (45-64), effects diverged: the HRA modestly over-represented men with asthma by 23% (OR=1.23, 95% CI 1.16-1.31) but there was no difference for women (OR=0.99, 05% CI 0.96-1.02). In contrast, for the oldest age group, 65 and over, there was no significant variance in the HRA population for males (OR=0.99, 95% CI 0.90-1.10) but there was a small over-representation of females (OR=1.16, 95% CI 1.13-1.18).

Figure 16: Comparison of asthma in the CCHS and HRA by age group and gender

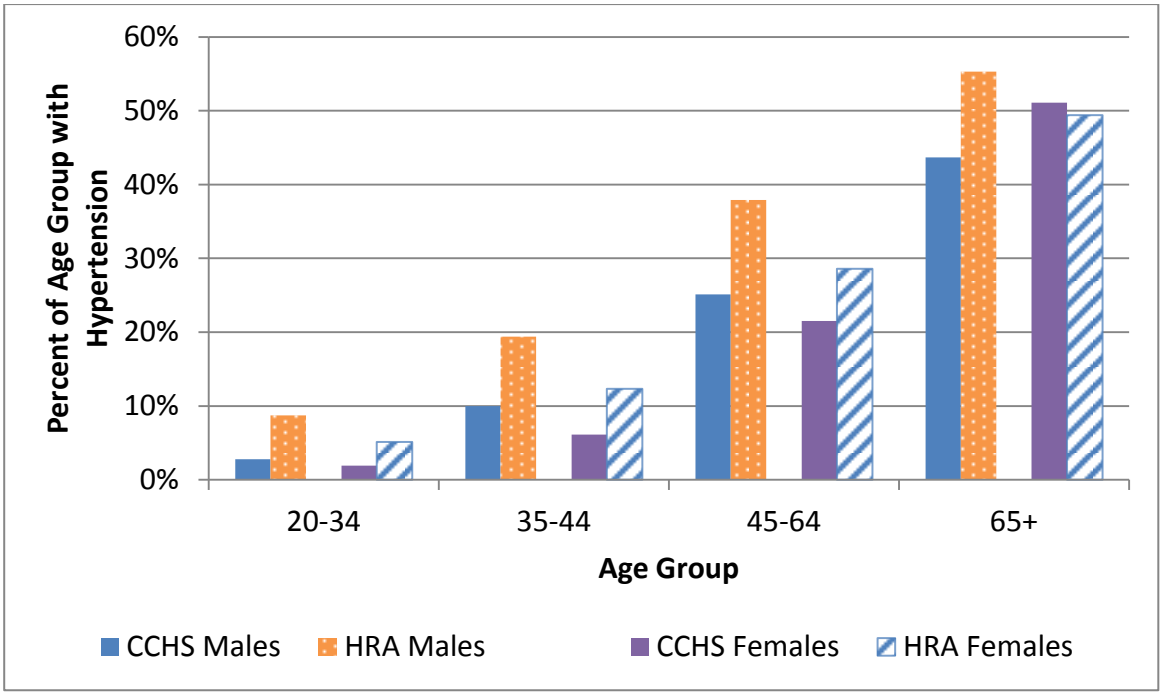


Hypertension

Hypertension is a primary risk factor for CVD and the prevalence increases with age (354). Thus, it would be expected that a heart health assessment might attract a disproportionately large number of people with this condition.

Figure 17 shows that for most age groups the anticipated trend occurred. Among those 20-34 years of age, odds were three times higher for HRA compared to CCHS respondents to report hypertension (for males OR=3.55, 95% CI 3.25-3.88; for females OR=2.97, 95% CI 2.78-3.21). These effects would be considered medium-sized if Cohen's criteria were applied (300) but large if Olivieri's suggestions were followed (312). Likewise, for those aged 35-44 HRA respondents had twice the odds of reporting hypertension (for males OR=2.14, 95% CI 2.01-2.27 and for females OR=2.11, 95% CI 2.00-2.21). Hypertension was also over-represented in HRA respondents 45-64 (for males OR=1.91, 95% CI 1.86-1.97 and for females OR=1.51, 95% CI 1.48-1.55). For seniors aged 65 and over, trends varied by gender. The HRA modestly over-represented senior males with hypertension (OR=1.77, 95% CI 1.74-1.81) but neither over- nor under-represented females with the condition (OR=1.02, 95% CI 0.98-1.06).

Figure 17: Comparison of hypertension in the CCHS and HRA by age group and gender

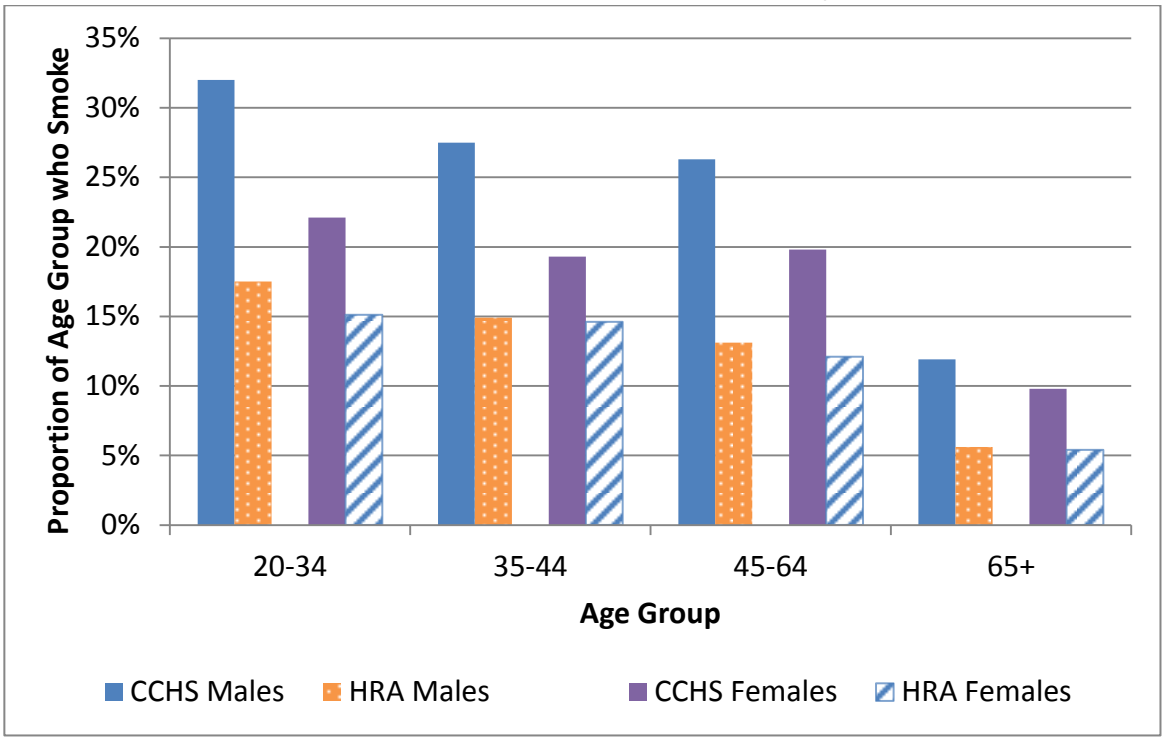


Smoking

Smoking is a major CVD risk factor (354). It is possible that those who smoke may be concerned about their cardiovascular health and thus utilize a health etool such as the HRA. It should be noted that there is some divergence between the two data sources in how this question was asked.

Figure 18 shows the report of occasional or daily smoking in the CCHS to a report of smoking (frequency not asked) in the HRA by age group and gender. It shows that for both men and women and for all age groups, smoking was consistently under-represented in the HRA sample. As shown in Table 4 in Appendix 4, depending upon the age group, the odds of being a smoker in the HRA were between 30% and 55% lower. However, these did not meet the RMPE recommended by Ferguson (313).

Figure 18: Comparison of smoking in the CCHS and HRA by age group and gender



Overweight or obesity

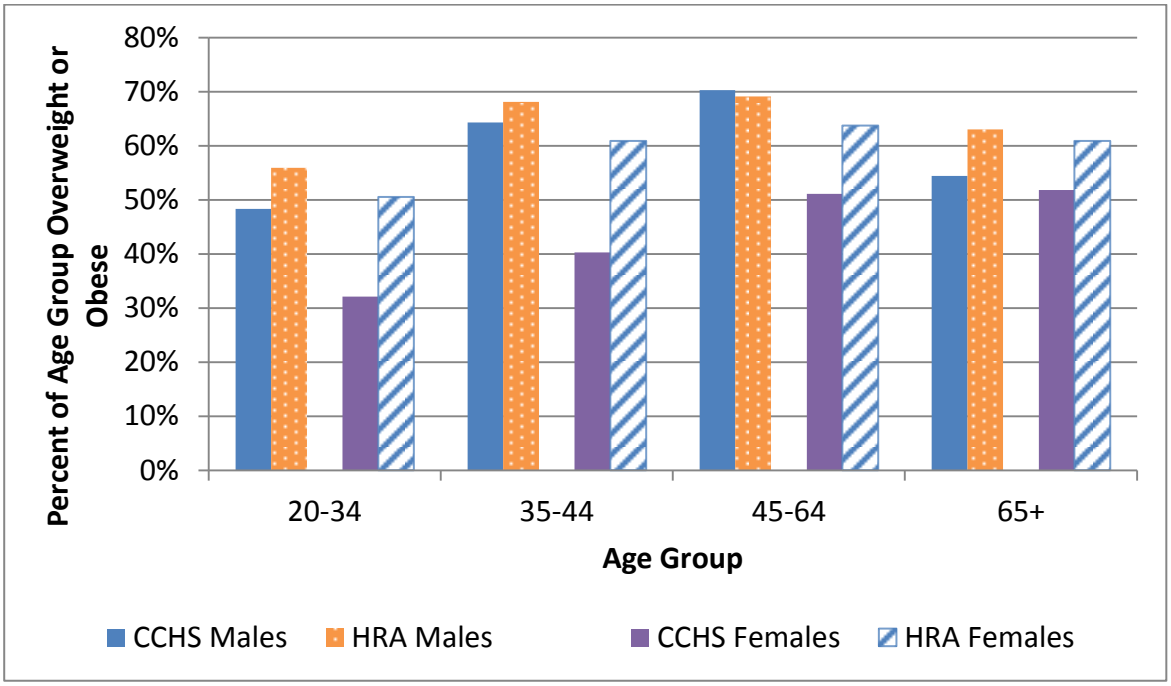
For both the CCHS and the HRA, BMI is calculated from self-reported height and weight. Overweight or obesity is considered a risk factor for CVD (354). However, it should be noted that during the study period it was mandatory for those registering for the HWAP to complete the HRA. As a result, it was anticipated that the HRA population would over-represent people with this condition.

Rates of obesity/overweight in the CCHS and HRA populations by age group and gender are shown in Figure 19. As anticipated this condition was consistently over-represented in the HRA sample. For women, the odds of being overweight in the HRA sample declined with age, being almost three-fold for those aged 20-34 or aged 34-44 years (respectively, OR=2.72, 95% CI 2.63-2.80 and OR=2.50, 95% CI 2.42-2.59), and about double for those 45-64 and those aged 65 and over (respectively, OR=1.96, 95% CI 1.92-3.00 and OR=1.88, 95% CI 1.80-1.96).

For males, over-representation of overweight and obesity was modest and greater for the youngest age group (OR=1.60, 95% CI 1.52-1.68), as well as the oldest age group (OR=1.43, 95% CI 1.36-1.50). For the 35-44 and 45-64 age groups, there were statistically significant (p<.001) but small increased odds that HRA males would be overweight/obese compared to the CCHS population (respectively, OR=1.19, 95% CI 1.13-12.6 and OR=1.11, 95% CI 1.08-1.15).

When all ages were combined, there was a modest effect for overweight/obesity to be over-represented in the HRA population for women (OR=1.51, 95% CI 1.49-1.52). For men, the effect was even smaller (OR=1.13, 95% CI 1.12-1.15).

Figure 19: Comparison of overweight/obesity in the CCHS and HRA by age group and gender



Mood disorder

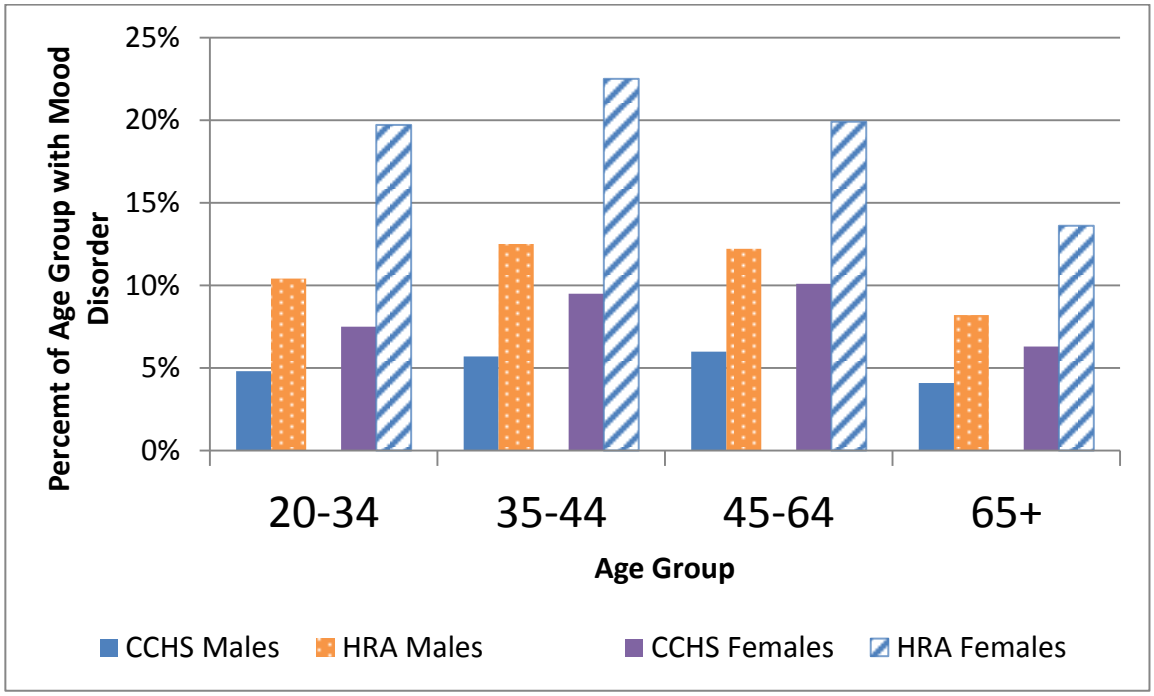
For some time, it has been observed that the general public considers stress a CVD risk factor (357, 358). In addition, medical research has reported that there is a complex association between mood disorders and CVD, with depression being reported as both a precursor and a sequella of heart disease (359, 360).

In the HRA, both depression and anxiety are captured in one question and qualified with the term "mood disorder." In the CCHS there are two questions: "Do you have a mood disorder such as depression, bipolar disorder, mania or dysthymia?" and "Do you have an anxiety disorder such as phobia, obsessive-compulsive disorder or a panic disorder?" For this comparison, the HRA category of mood disorder was compared to a prevalence reflecting the total number of individuals responding yes to either or both of the CCHS questions. Given the greater number of conditions named in the CCHS questionnaire, it might be expected it would capture more individuals self-reporting these problems.

The proportions by age group and gender who report mood disorders in the CCHS and HRA populations are shown in Figure 20. As it shows, in both populations the

prevalence tended to be higher in women than men and decreased somewhat in the oldest age group. This pattern reflected population trends reported in the epidemiologic literature (361).

Figure 20: Comparison of mood disorders in the CCHS and HRA by age group and gender



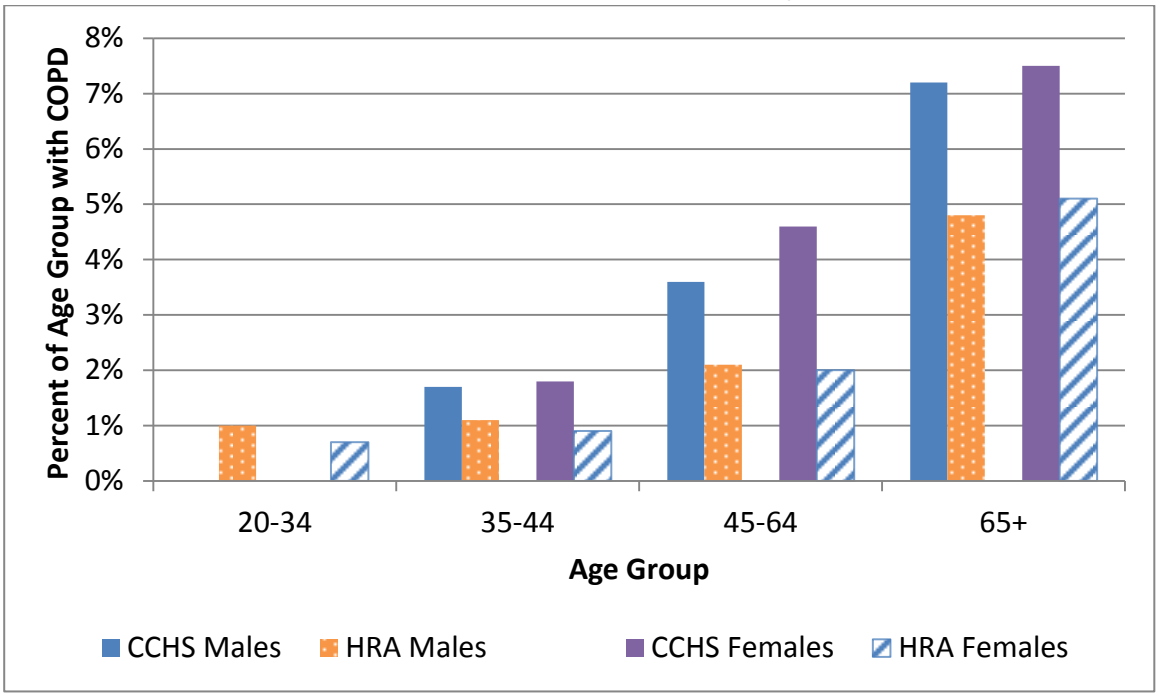
For men of all ages, the odds of reporting a mood disorder in the HRA were twice that of the CCHS (overall OR=2.21, 95% CI 2.14-2.28). Odds ranged from a high of 2.49 (95% CI 2.29-2.70) for the 20-34 age group to a low of 2.21 (95% CI 2.03-2.43) for the 65 and over group.

For women, the odds or reporting mood disorders in the HRA varied by age. The lowest OR was for the 65 and over group: OR=2.44 (95% CI 2.30-2.59), a value that meets the RMPE (313). Odds were higher for the 45-64 group: OR=6.44, 95% CI 6.28-6.60, a medium-size effect according to Cohen (300) but large according to Ferguson (313) or Olivier (312). For the 20-34 and 35-44 age groups, the HRA over-represented the prevalence of mood disorders three-fold (respectively, OR=3.21, 95% CI 2.08-3.33 and OR=2.71, 95% CI 2.60-2.81).

COPD

Like arthritis and asthma, there is no direct relationship between COPD and CVD. Thus, it would not be expected that people with COPD would utilize an etool focusing on cardiovascular health.

Figure 21: Comparison of COPD in the CCHS and HRA by age group and gender



Due to small numbers, the CCHS could not produce reliable estimates for the prevalence of COPD among men or women age 20-34 years. For all other age groups, there was a consistent trend for the HRA to under-represent COPD (see Figure 21). For men, the odds were between 30% lower (for seniors aged 65 and over, OR=0.70, 95% CI 0.62-0.78) to 40% lower (for age 35-44, OR=0.59, 95% CI 0.54-0.65). For women, the odds ranged from 30% lower for the 65 and over group (OR=0.70, 95% CI 0.64-0.77) to 56% lower (for the 35-44 group OR=0.44, 95% VI 0.41-0.47). These would be considered small affects according to Cohen's cut-offs.

What does this comparison suggest?

Comparing Canadian HRA users to the general population of Canada showed there were frequently substantive difference by gender, age, socioeconomic status as represented by level of education, at least one health behaviour (smoking), and several chronic conditions. In other words, the HRA population was not representative of the general population. Non-CVD-associated conditions such as asthma and COPD varied only modestly from the CCHS estimates whereas diabetes and hypertension, which are CVD risk factors, were over-represented in the HRA. These results suggest the HRA may be effective in reaching Canadians at increased risk of CVD because of diabetes or hypertension, particularly younger adults with diabetes.

Mood disorder was an interesting exception. Despite the fact that the HRA question provided fewer prompts in the form of names of specific disorders, mood

disorder was over-represented among males and females. Possible reasons for this finding are unknown and the trend may warrant further research to determine if the difference is real or an artefact of differences between the two survey questions.

Does an incentive change HRA users?

As described in Chapter 3, during the data collection period, the HSF offered a promotion in which Air Miles customers were sent solicitation emails and offered a relatively modest, non-monetary incentive of ten bonus points for completing the HRA and another ten if they enrolled for the eSupport email service. The HSF's objective in undertaking the Air Miles promotion was to increase the reach of the program, particularly among what it suspected were under-represented segments of the population, such as younger males.

During the study period, 72,454 or 60.1% of records were created by users coming through the Air Miles promotion. Although the promotion had a strong effect on the number of people completing the HRA, there is limited evidence that it changed the type of users (see Table 5 in Appendix 4). The Air Miles promotion had no significant effect on the proportions of users by gender (32.1% of Air Miles and 31.9% of non-Air Miles users were male, Cramer's V_{1df} =.003, *p*=.354) and only a small effect on mean age (48.4 [14.1] vs. 48.8 [14.1], Cohen's *d*=.028) or age groups (*eta*=.043). The effect of Air Miles status on the distribution of participants was also small for education (Cramer's V_{1df}=.026), employment status (Cramer's V_{1df}=.073) and type of work (Cramer's V_{1df}=.063).

Air Miles participants had a lower mean number of vascular diseases (0.5 [0.9] *vs.* 0.7 [1.0], Cohen's *d*=.210), modifiable risk factors (2.5 [1.4] *vs.* 2.7 [1.4], Cohen's *d*=.143) and non-modifiable risk factors (2.0 [1.4] *vs.* 2.2 [1.5], Cohen's *d*=.138). As a result, Air Miles participants had a slightly lower number of total CVD risk factors (5.0 [2.4] *vs.* 5.6 [2.5], Cohen's *d*=.245). However, effect sizes for all comparisons were small (i.e., < 0.60, which is the cut-off for a medium-sized effect).

As shown in Table 5 (Appendix 4), there appeared to be a trend for the Air Miles participants to be healthier than their non-Air Miles counterparts but for non-modifiable and modifiable risk factors and report of vascular diseases none of the comparisons were associated with even a medium effect size. There was a small-to-medium-sized effect for Air Miles participants with hypertension to report good hypertension control (72.8% vs. 49.9% reported their blood pressure was in a healthy range "most of the time," Cramer's V_{1df}=.254). But there was no difference in blood glucose or lipids control or medication adherence.

What does this comparison suggest?

It appears the Air Miles incentive had only small effects on the type of people completing the HRA. It is possible the incentive was insufficient to change behaviour and to get those who are not health-oriented or not comfortable with technology to visit the site (85). For example, Khadjesari *et al.* found an incentive had to be worth at least £10 before it influenced follow-up rates in the more controlled setting of an RCT (362), while Alexander *et al.* estimated the average cost of recruiting an etool trial participant to be $32(US) per person and the cost of retention $70(US) (112). The relatively modest (10 bonus points) and non-monetary nature of the incentive may have been insufficient to change the behaviour of those not interested in, or even resistant to, health information, such as 30% who Wilkins and Navarro described as placing a low value on maintaining or improving their health (96).

Is the HRA population similar to samples recruited for online health etool RCTs?

Harle has argued that studies of operating health promotion programs have strong ecological validity as they operate in the real world, rather than the artificial environment of experimental research (229). As described in Chapter 1, much of the research to date on health promotion etools has been conducted using experimental paradigms such as RCTs which utilize participant inclusion and exclusion criteria and recruit using specific methods or among select populations or at specific sites. In many cases such research samples have reflected attributes of health information seekers, such as being largely female, health conscious and more highly educated (48, 49, 58, 62, 74), similar to the HRA population. So could samples created for RCTs be representative of users of freely-available health etools? To explore this issue, the HRA population was compared to the samples generated for three RCT online health etools similar, albeit not identical, to the HRA.

1) A Swiss physical activity etool

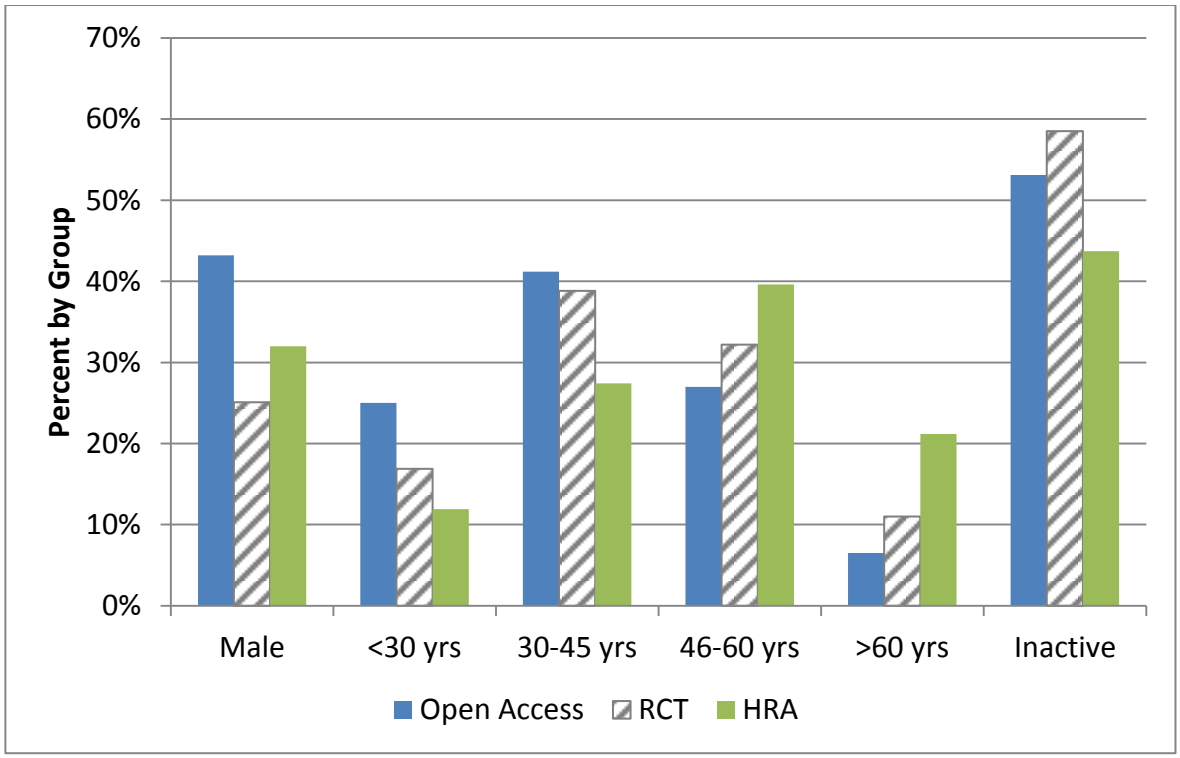
In 2010, Wanner *et al.* compared participants of a RCT (n=836) of a Swiss physical activity etool, *Active-online*, with open access users of the same tool (n=5,083) (308). For open-access users who registered with the program, three data points were captured: gender, age, and whether the individual met the health-enhancing physical activity (HEPA) recommendations of $\geq$30 minutes of moderate-intensity activity on five or

more days per week or ≥20 minutes of vigorous intensity activity on three or more days. The HEPA cut-off may be compared to the HRA physical activity question.

In the Wanner *et al.* study, mean age varied significantly between the two arms: 39.1, 95% CI 39.0-39.2, years for open access participants *vs.* 43.1, 95% CI 42.2-44.0, years for the RCT arm (p<.001) (308). As well, there was a significant difference by gender, with 55.1% of open-access and 74.9% of RCT participants being female (308). The other significant difference concerned retention: despite email reminders for registered open-access users, attrition was higher than among trial participants (308).

Figure 22 compares the proportions of the three samples (open-access, RCT and the HRA) by gender, age group, and physical inactivity (308). As it shows, compared to the open-access arm, HRA users were less likely to be male (OR=0.62, 95% CI 0.59-0.66, representing a 1.5-fold reduced risk), less than 30 years of age (OR=0.41, 95% CI 0.38-0.43, representing a 2.44-fold difference), 30 to 45 years of age (OR=0.53, 95% CI 0.50-0.56, a 1.89-fold difference), and inactive (OR=0.69, 95% CI 0.65-0.73, a 1.45-fold reduced likelihood). HRA participants were more likely than the open access participants to be >60 years (OR=3.85, 95% CI 3.44-4.30) or 46-60 years (OR=1.77, 95% CI 1.67-1.89). Thus, even under the open access condition, differences occurred between the populations of the two etools.

Figure 22: Gender, age groups and activity status for Wanner et al. open access and RCT participants compared to HRA population



The HRA was then compared to the RCT arm. Those <30 and 30-45 years of age continued to be under-represented in the HRA (respectively, OR=0.66, 95% CI 0.55-0.80 and OR=0.60, 95% CI 0.52-0.68), while those 46-60 and >60 were over-represented (for 46-60, OR=1.38, 95% CI 1.19-1.60 and for >60 OR=2.17, 95% CI 1.75-2.70).

When compared to the open-access arm, males were under-represented in the HRA but when compared to the RCT arm they were over-represented (OR=1.41, 95% CI 1.20-1.64). There was also a difference for physical inactivity. Whereas the HRA had a 44% likelihood of under-representing inactive participants when compared to the open-access arm, when compared to the RCT there was no significant difference (OR=1.03, 95% CI=0.91-1.16, p=0.663).

In summary, Wanner ***et al.'s*** study found that there were significant difference between open-access and RCT participants for the same etool. Moreover, the HRA population varied in some significant ways from both of Wanner *et al.*'s two arms. Some but probably not all of this difference could be due to the type of etool and their country of origin.

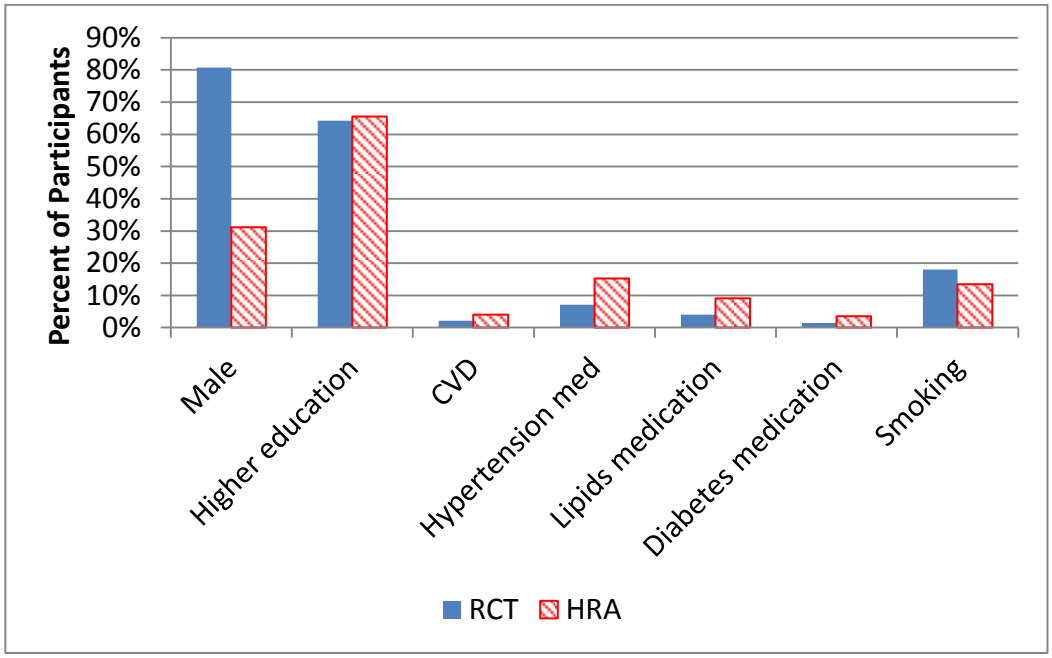
2) A Dutch workplace CVD risk assessment

In 2011, Colkesen *et al.* published a report on a web-based health risk assessment similar to the HRA in that is focused on CVD risk and generated tailored health advice (309). This etool was not freely available but marketed to employees at a single Dutch worksite, of which 772 volunteered to participate. Four variables appeared to be similar between the HRA and the Dutch study: gender, current smoking, being prescribed anti-hypertensives, and being prescribed diabetes medication. An additional three variables were similar: higher education (in the RCT defined only as "high" compared to "low" or "medium"; those in the "high" category may be comparable to the HRA category of "college or university graduate"), being prescribed medication for dyslipidemia (in the RCT specified as statins), and personal history of CVD (as this was not defined, the RCT proportion was compared to the HRA proportion reporting heart disease and/or stroke/TIA).

To more closely match the RCT sample, for this analysis the HRA database was limited to those employed full- or part-time (n=69,280 or 57.5% of the entire HRA population). Figure 23 shows the proportions for the seven variables. ORs met the RMPE for five of the variables: male gender (OR=0.01, 95% CI 0.01-0.02), CVD (OR=1.96, 95% CI 1.19-3.23), hypertension medication (OR=2.35, 95% CI 1.78-3.09), medication for dyslipidemia (OR=2.40, 95% CI 1.67-3.44), and medication for diabetes (OR=2.50, 95% CI 1.37-4.53, *p*=.002). There were no significant differences between the two samples in the

proportions reporting higher education (OR=1.06, 95% CI=0.92-1.22, *p*=0.540) and only a modest difference (29% reduced likelihood) for smoking (OR=0.71, 95% CI 0.59-0.85). Thus, even when the etool concerns the same topic (CVD) and the population is restricted to the same sub-set (those employed), the sample recruited for an RCT differed in several ways from the HRA population.

Figure 23: Comparison of worksite RCT population (Colkesen et al.) and working HRA population



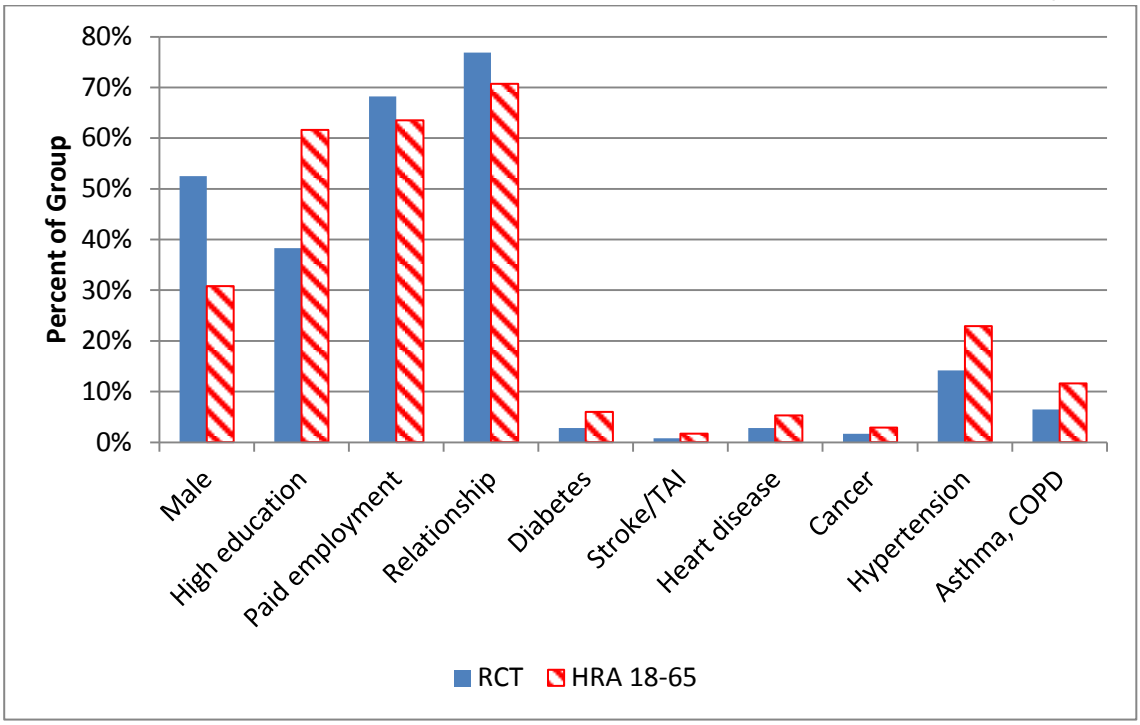
3) A Dutch lifestyle intervention marketed to the general population 18-65 years

In Schulz *et al.*'s RCT of a Dutch web-based tailored lifestyle intervention, three study groups were recruited among the general population in two provinces (310). To be included, participants had to have a valid email address, a computer with Internet access and basic Internet literacy, and be between 18 and 65 years of age (310). As shown by Schulz *et al.*, there were no significant differences between the two experimental (n=552 and 517) and one control group (n=664) in demographics or diseases (310). Thus, total proportions from the RCT (n=1,733) were compared to HRA users aged 18 to 65 (n=107,358 or 89.1% of the entire HRA population).

Figure 24 compares proportions from the Schulz *et al.* study to those in the HRA. Although the difference between the two populations was minimal for employment status (OR=0.80, 95% CI 0.73-0.90) and those married or in a relationship (OR=0.73, 95% CI 0.65-0.81), four of the ten variables met the RMPE: male gender (OR=0.40, 95% CI 0.37-0.44), higher education (OR=2.58 95% CI 2.34-2.85), diabetes (OR=2.25, 95% CI 1.69-3.00), and stroke/TIA (OR=2.27, 95% CI 1.32-3.93). Moreover, the OR approached the RMPE for the combined category of asthma and COPD (OR=1.88, 95% CI 1.55-2.27),

hypertension (OR=1.79, 95% CI 1.56-2.05) and cancer (OR=1.70, 95% CI 1.18-2.45). The only variable for which there was no significant difference was report of heart disease (for the RCT, a combination of the categories “heart attack” and “other serious heart diseases”): OR = 1.18, 95% CI 0.89-1.57, *p*=.248.

Figure 24: Comparison of RCT sample (Schulz et al.) to HRA sample (18-65 years)



What do these comparisons suggest?

Although there were some similarities between the HRA population and three RCT samples (which probably reflects the fact that both are drawn predominantly from those segments of the population that are Internet health information seekers), there were several substantive differences as well. It is not surprising differences should be observed between a population obtained through secondary analysis of observational data and samples generated for experimental research. Although the HRA population and RCT samples self-select, their motives and context for participation vary. The HRA database consists of information submitted voluntarily by individuals for the purpose of receiving personalized feedback about their own health status (a form of self-interest), whether serious in intent or for the purpose of entertainment. The HRA is not positioned as a research study, there are no strict inclusion or exclusion criteria to meet, no up-front requirement to give informed consent in order to access the tool, and there is no need for ongoing participation. Consent for the use of data for research purposes is asked, but only after the user has completed the questionnaire. In contrast, although people are generally

more likely to volunteer for studies that are personally relevant (363), subjects who enroll in RCTs do so knowing they are participating in an intervention with the objective of benefiting research rather than themselves. Moreover, unlike HRA users, RCT subjects must meet inclusion/exclusion criteria, give informed consent prior to enrollment, and may be asked to commit to extended participation (e.g., follow-ups, either online or involving travel).

In research, it is recognized that volunteers are likely to differ from non-volunteers in several demographic, health and/or psychological characteristics (363, 364). For example, as described by Golomb *et al.*, poor health may be a barrier to volunteering for research studies and the effect appears to increase with age (365). Although conducting a study through the Internet may mediate this effect by removing the requirement to travel to research centres, it may not eliminate psychological or social barriers. Indeed, as discussed by Eysenbach and Wyatt, research samples recruited online suffer from not only the volunteer effect but the bias introduced by the non-representative nature of Internet users (366). Whether research is conducted on- or off-line, different sources of referral and volunteer bias must be considered when appraising experimental research (366, 367).

Although there were some similarities between the HRA population and samples recruited for the convenience sample of three RCTs, there were also differences. It is possible, for example, that the reluctance for less healthy adults to participate in research (365) may help to explain why, compared to the HRA population, the RCT samples were more likely to be physically active (308) and less likely to report chronic conditions such as hypertension or diabetes (309, 310). However, there may be several, and perhaps even different, types of biases affecting RCT and open-access samples.

In summary, this review suggests attempts to generalize results from samples created for RCTs to the populations that use freely-available etools for self-assessment must be viewed with caution. Until the digital divide is erased, online populations will continue to differ in some ways from the wider populations from which they are drawn (366). Finally, although the Internet itself is borderless, applying results from RCTs conducted in one country may not be appropriate if an etool is based in another setting.

Summary

Initial analysis of the HRA population presented in Chapter 5 showed strong trends by gender, age, level of education and other demographics in ways that resembled previous research on Internet health information seekers (48, 348). Such people have

been described as more health-oriented or health conscious, often because of health concerns (79, 81, 84, 348). However, although the HRA appeared skewed, the extent to which there was a systematic difference from the general population or from the sort of samples commonly recruited in RCTs was unknown.

First, the question of whether the HR reflects the general population of Canada was addressed. This analysis showed Canadian HRA participants were significantly:

- more likely to be female, between 45 to 64 years of age and to have graduated with a college or university education;
- less likely to be smokers and, perhaps because of the presence of the HWAP, more likely to report being overweight or obese;
- more likely to report conditions such as hypertension, asthma and mood disorders and less likely to report COPD; as well, there was a trend for the HRA to over-represent younger users with diabetes or arthritis but to under-represent older adults with these conditions.

Furthermore, when education was weighted by age and gender, differences between the HRA and the general population persisted. This suggests a systemic bias in the type of Canadians attracted to the HRA. This is not surprising, as people who utilize a freely-available health etool are not only health information seekers (i.e., more interested in health messaging and more active in searching for it) but those who have selected a particular medium (i.e., the Internet) and topic (cardiovascular health) (13, 51).

The issue of Internet access is a key consideration. Although those who are older and of lower SES may be at increased risk of CVD, they also belong to groups less likely to have Internet access. The 2012 Canadian Internet Use Survey, for example, reported that only 28% of Canadians 65 and over use the Internet, compared to 95% of those 16-24 years of age; likewise, Internet use was 62% for those in the lowest income quartile but 95% in the highest (54). Thus, people who utilize the HRA may reflect those in the population that self-select to be health information seekers and also have the means to do so through the Internet.

The second question concerned the potential of an incentive to change the type of people who complete an open-access health etool. For this analysis, the natural experiment afforded by the HSF's Air Miles promotion was utilized. Analysis showed that although the promotion increased the number of people who competed the HRA it did not significantly alter their demographic or health profiles. It is possible the relatively modest and non-monetary nature of the Air Miles incentive was insufficient to change the

behaviour of people who are not inclined to be Internet health information seekers. Further research with larger or different types of incentives may be needed.

Finally, the third question addressed in this chapter concerned the generalizability of RCT samples to open-access, non-experimental populations. In some respects, Internet health information seekers and participants in etool RCTs share many characteristics (117, 368, 369), which may a similar bias towards more health literature and health information seeker populations. But does that mean results from RCTs can necessarily be generalized to freely-available etools and populations? To answer this question, the HRA population was compared to samples recruited for three etool RCTs that were somewhat similar to the Heart&Stroke Risk Assessment (e.g., concerned CVD or modifiable CVD risk factors). Comparisons showed that the RCT samples recruited by Wanner *et al. (308),* Colkesen *et al.* (309) and Schulz *et al.* (310) differed in substantive ways from the HRA population in gender (308-310), age (308), physical activity (308), and report of chronic conditions (309, 310). Differences persisted even when the HRA sample was limited to resemble the inclusion criteria of the different RCTs.

RCTs of health etools draw upon the same population as open-access health etools: those who have Internet access and are receptive to health issues (i.e., may be health conscious). However, it is not surprising that there are also differences between those who are looking for information for themselves (HRA users) and those who are willing to participate in a research study (RCT participants). Not only are there differences in their motivation for using an etool, compared to open-access users, RCT samples must, by their very nature, be effected by volunteer bias (363, 365). The extent to which RCT samples resemble open-access users is unclear and caution must be used in generalizing from one to the other. In the case of the HRA, differences were found but the HRA is, after all, only one open-access health etool.

Given the paucity of published research on open-access etools, it can be argued the nature of user populations remains poorly understood. For example, even though there has been a recent publication concerning users of the open-access Heart Age calculator, the only sociodemographic information collected, and therefore reported, for this etool was age and gender (232). For those variables, even though both Heart Age and the HRA concern CVD and are open-access, there are significant differences: mean age for Heart Age participants was 42.9 (14.0) years compared to 48.6 (14.1) for the HRA and males constituted 44.0% of the Heart Age sample but were only 32.0% of the HRA population (232). Whether the two populations varied by education, employment, or other variables cannot be studied. Until more open-access etools share demographic and health information about their users, it is impossible to determine whether various user

populations are similar or differ according to the health issue being addressed (e.g., chronic *vs.* acute conditions or between different types of chronic diseases), country of origin, or type of intervention. Understanding the characteristics of open-access etools could in turn contribute to conducting and interpreting comparisons of specific etools to RCT samples. It is possible, for example, that variance between open access etool populations is equal to or even greater than variance between open access and RCT samples.

7. Segmentation

Analysis demonstrated the HRA population differs from the general population of Canada, as well as samples assembled for experimental studies of health etools. However, regarding the HRA population as monolithic would limit our understanding. In this chapter, different segmentation procedures and clustering variables were used to determine if meaningful and useful groups could be created within the HRA population.

Correlation matrix and identification of clustering variables

Most segmentation procedures require at least moderate correlation between variables although, as discussed, excessive collinearity should be avoided. Table 6 shows the correlation between the recoded variables created for analysis. Because lifestyle healthiness score incorporates information on the prevalence and stage of change for modifiable risk factors, it is not surprising that there were strong negative relationships between it and the number of modifiable risk factors (*r*=-.885) and total number of health concerns (*r*=-.548). Likewise, as total number of health concerns is a count of number of reported vascular diseases and modifiable and non-modifiable risk factors, moderate to strong correlation between these factors was also anticipated and observed (.551 for number of vascular diseases, .730 for number of nonmodifiable risk factors, and .634 for number of modifiable risk factors).

Table 6: Correlation (Pearson's r) between constructed variables

Variable	Number of Vascular Diseases	Number of Non-modifiable Risk Factors	Number of Modifiable Risk Factors	Total Number of Health Concerns	Overall Lifestyle Healthiness Score
Number Vascular Diseases		.229	.050	.551	-.028
Number Non-modifiable Risk Factors	.229		.091	.730	-.068
Number Modifiable Risk Factors	.050	.091		.634	-.885
Total Number of Health Concerns	.551	.730	.634		-.548
Overall Lifestyle Healthiness	-.028	-.068	-.885	-.548	

For all correlation co-efficients, p<.001

In the general population, the presence of modifiable risk factors is typically associated with an increased incidence or prevalence of chronic diseases, particularly vascular-related diseases (hence, the labeling of behaviours such as smoking as "risk factors"). However, in the HRA population the relationship between the modifiable risk factors and vascular diseases may be statistically significant as indicated by the *p* value but is weak (r=.050). The relationship between number of vascular diseases and number of non-modifiable risk factors is stronger (r=.229), and thus perhaps a more appropriate combination for segmentation.

Table 6 also illustrates the challenge of *p* values in the analysis of large databases. Although all correlations were statistically significant, several, such as those between number of modifiable and non-modifiable risk factors (*r*=.091) or between lifestyle healthiness and number of vascular diseases (*r*=-.028) and non-modifiable risk factors (*r*=-.068), were small. Such correlations may be statistically significant but offer little meaningful information for understanding the relationship between variables.

Selection of clustering variables

With a data base as large as the HRA, numerous combinations of variables could be used for clustering purposes. For this study, analysis was limited to five combinations:

1. Number of vascular diseases, modifiable risk factors, and non-modifiable risk factors;

2. Number of health concerns and lifestyle healthiness score;

3. Age in years, lifestyle healthiness score, number of vascular diseases and number of non-modifiable risk factors (to test the effect of using age as a clustering variable);

4. Modifiable and non-modifiable risk factors as binary (present/absent) nominal variables; and

5. Modifiable risk factors and vascular diseases as binary (present/absent) nominal variables.

Each combination of variables was subjected to as many of the three types of segmentation procedures (k-means, two-step, and latent class analysis) possible given the type of clustering variable (nominal or interval). This made it possible to compare results not only between different clustering variables but also between different segmentation procedures.

Approach 1: Number of vascular diseases and modifiable and nonmodifiable risk factors

Because vascular diseases and modifiable and nonmodifiable risk factors are the primary components of CVD risk, they would appear to be logical factors with which to segment the HRA population. To test the ability of these factors to form meaningful groups, all three were submitted to LCA and k-means clustering and two-step clustering. A statistically significant solution could not be achieved using LCA (see Table 1 in Appendix 5) but convergence was obtained for the two forms of cluster analysis.

When k-means clustering was attempted, there was convergence for three-, four- and five-group solutions. The four-group solution, K-means Solution 1, had good internal consistency, as convergence was achieved even when the file was split by gender or Air Miles status.

Group sizes were quite equitable: 29.0%, 28.0%, 24.8% and 18.2% of cases. Group membership had a medium-sized effect for only one variable not used for clustering: age in years (ϖ=.127). Effect sizes for other non-clustering variables were small (Cramer's V_{1df}<.30 or ϖ<.06, see Table 2 in Appendix 5), including readiness to change modifiable risk factors (for all, Cramer's $V_{1df}\leq$.197 indicating small effects).

Groups varied significantly in the distance of cases from cluster centres (ϖ=.30). Of the three clustering variables, effect size for group membership was strongest for modifiable risk factors (ϖ=.825), followed by non-modifiable risk factors (ϖ=.747) and number of vascular diseases (ϖ=.339); these findings may reflect the prevalence of these factors in the HRA population.

Figure 25 shows the proportions reporting modifiable risk factors and vascular conditions by K-means Solution 1 group. In contrast to what would be expected from the trends shown in Figure 9, the groups with the older median ages did not have higher rates of vascular conditions.

Figure 25: Proportions by K-means Solution 1 group for modifiable risk factors and vascular conditions

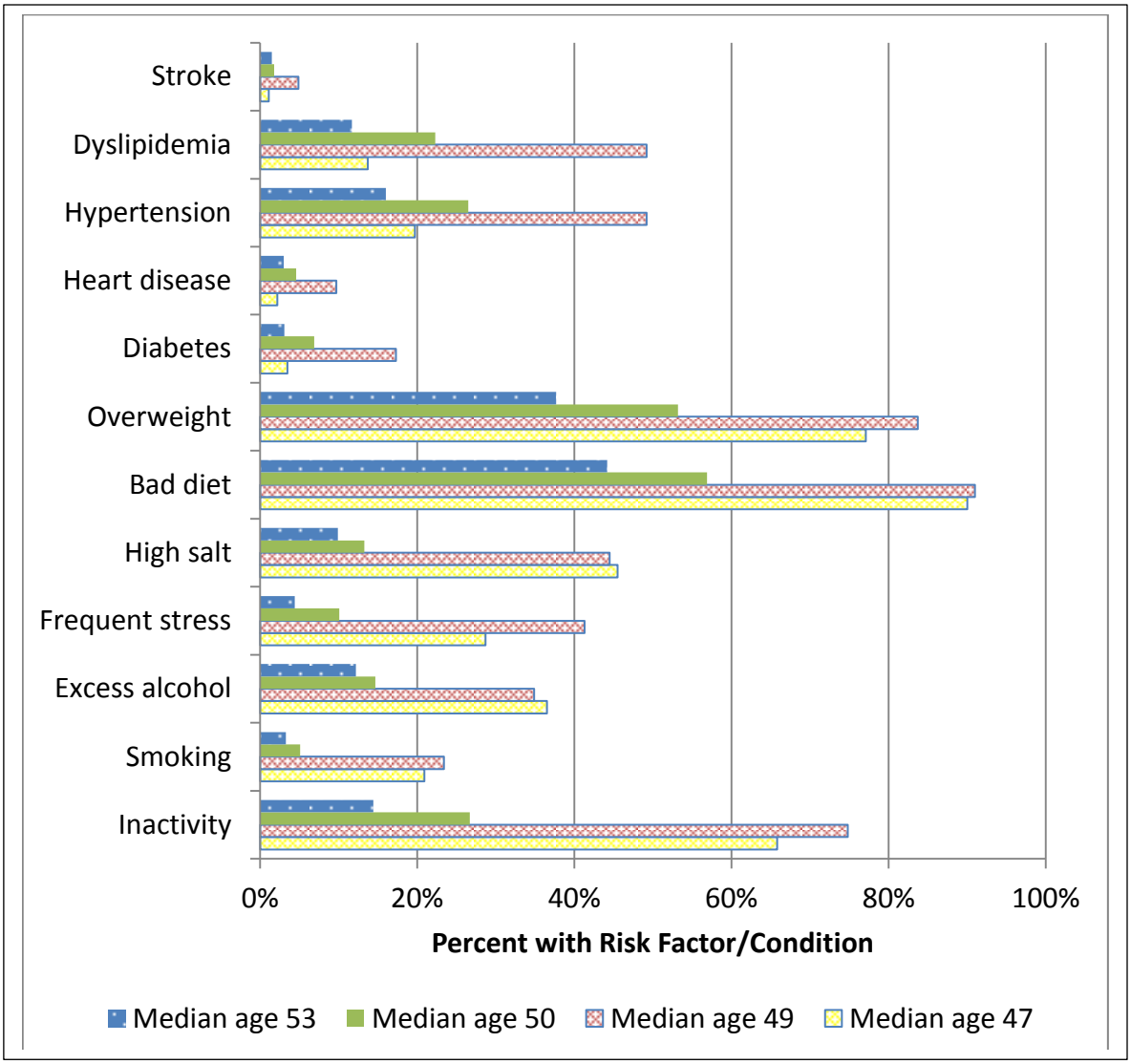


Table 7 shows means for the number of vascular diseases, modifiable risk factors, non-modifiable risk factors and other interval data for K-means Solution 1, as well as Two-step Solution 1. Means did not support the sort of age-related trends suggested by Figure 9. In summary, there was evidence suggesting the solution had poor face validity.

Table 7: Comparison of means between K-means and Two-step Solutions 1

	Group A	Group B	Group C	Group D
K-means Solution 1				
Mean age (years)	46.4	47.8	48.8	51.3
Median age (years)	47	49	50	53
Mean number vascular diseases	0.4	1.3	0.6	0.4
Mean number modifiable risk factors	1.4	3.9	1.8	1.3
Mean number non-modifiable risk factors	1.1	3.6	3.2	0.8
Number of health concerns	5.1	8.8	5.6	2.4
Lifestyle healthiness score	26.2	25.6	30.9	32.2
Proportion of population (%)	29.0%	18.2%	28.2%	28.1%

	Group A	Group B	Group C	Group D
Two-step Solution 1				
Mean age (years)	44.3	48.9	49.3	58.6
Median age (years)	45	50	51	59
Mean number vascular diseases	0.3	0.4	0.1	2.5
Mean number modifiable risk factors	3.7	1.6	1.2	2.7
Mean number non-modifiable risk factors	2.0	3.0	0.7	2.8
Number of health concerns	6.0	5.1	2.1	8.0
Lifestyle healthiness score	26.0	31.3	32.3	28.7
Proportion of population (%)	41.7%	22.7%	20.5%	16.2%

When the two-step procedure was used, a two-group solution was generated with a silhouette co-efficient of 0.6, suggesting good cohesion and separation of groups. A four-group solution (Two-step Solution 1) could be forced, although the silhouette co-efficient declined to 0.4, suggesting only fair cohesion and separation. Group sizes were less equitable than was the case for K-means Solution 1, with 41.7% falling into one group, and the other three groups comprising 22.7%, 20.5% and 15.2% of cases.

Proportions, means and effect sizes for groups formed by Two-Step Solution 1 are provided in Table 3 in Appendix 5. For the clustering variables, effect sizes followed the same pattern as in K-means Solution 1, being largest for modifiable risk factors (ϖ=.843), followed by non-modifiable risk factors (ϖ=.547) and vascular diseases (ϖ=.340). Of variables not used for clustering, there was a large effect for age in years (ϖ=.340) and a medium-sized effect (Cramer's V_{3df}=.200) for age groups. Medication use had a medium-sized effect by cluster (Cramer's V_{1df}=.380) but as discussed earlier this variable is confounded by age.

Figure 26 shows proportions of those reporting modifiable risk factors and vascular conditions in Two-step Solution 1. For all modifiable risk factors except salt, the youngest age group (median age 45) had the highest proportions; however, proportions did not decrease by age, as might be expected from the trends in Figure 9. Instead, the oldest age group had the second-highest proportions, with the exception of salt for which it had the highest proportion.

For the vascular conditions, rates were highest for the oldest age group but did not decline in a linear fashion with the age of the groups. In other words, this solution did not conform to the age-related trends seen in Figure 9. These findings suggest that this solution had poor face validity and was not helpful in understanding the HRA population.

Figure 26: Proportions by Two-Step Solution 1 group for modifiable risk factors and vascular conditions

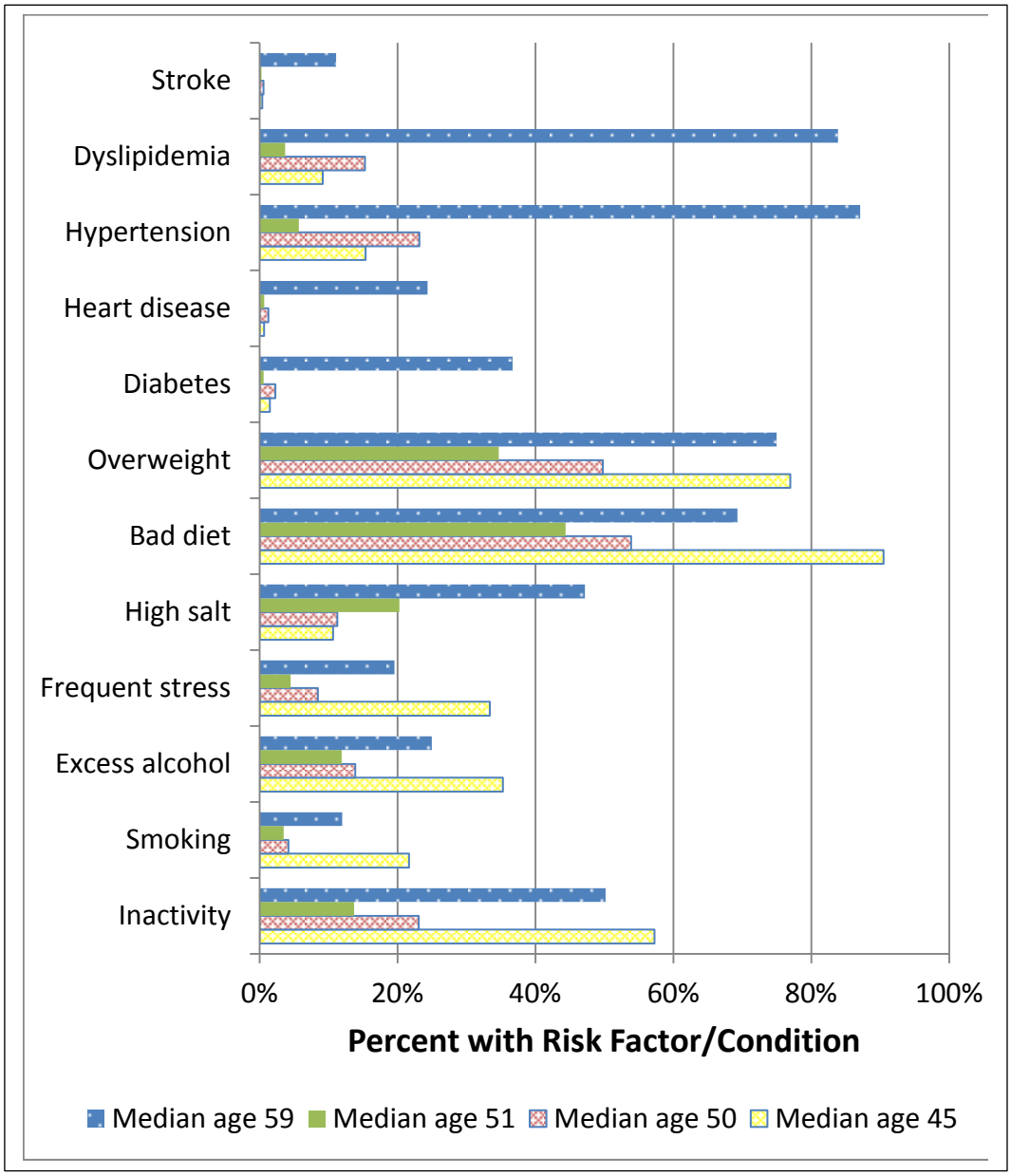


Table 8 summarizes all three segmentation procedures attempted using this set of clustering variables. Each of the two clustering solutions had good internal consistency but poor face validity. When groups were renumbered in order by age, agreement between the k-means and two-step solutions was below chance (61,227/118,941 = 51.5%, Cohen's *kappa*=.367, $p<.01$) suggesting the different procedures created quite different groups. More importantly, these solutions did not appear to be particularly helpful in expanding our understanding of the HRA population.

Table 8: Comparison of Segmentation Solutions: Number of Vascular Diseases and Modifiable and Non-Modifiable Risk Factors as Clustering Variables

	K-means Clustering (K-means Solution 1)	**Two-Step Clustering (Two-step Solution 1)**	**Latent Class Analysis**
Potential solutions	3-, 4- or 5-groups	System recommends 2-group solution; 4-group solution can be forced	None generated
Internal consistency	Good	Good	N/A
Group sizes of the selected solution	Equitable for 3- and 4-group solutions; less equitable for 5-group	For 4-group solutions, membership sizes range between 15% and 42%	N/A
Large-to-small group size ratio	1.59	2.74	N/A
Face validity for 4-group solution	Poor	Poor	N/A
Differentiation for 5-group solution	Medium-sized effect for age in years	Large-sized effect by age in years; medium-sized effect for age group and medication use	
Distance of cases from cluster centre	Large	N/A	N/A
Silhouette Coefficient (2-step only)	N/A	Fair (0.4)	N/A
Classification error rate (LCA only)	N/A	N/A	N/A
Agreement between 4-group solutions	51.5%, Cohen's *kappa*=0.367, *p*<.001		N/A

Approach 2: Clustering using number of health concerns and lifestyle healthiness scores

As previously shown in Table 6, total number of health concerns and overall healthiness scores had a Pearson's *r* of -0.548. This suggests a moderately strong negative relationship between the two variables but not excessive collinearity. These variables could be used to create k-means and two-step cluster solutions but were unable to generate a significant LCA solution (for more information on the latter, please refer to Table 4 in Appendix 5).

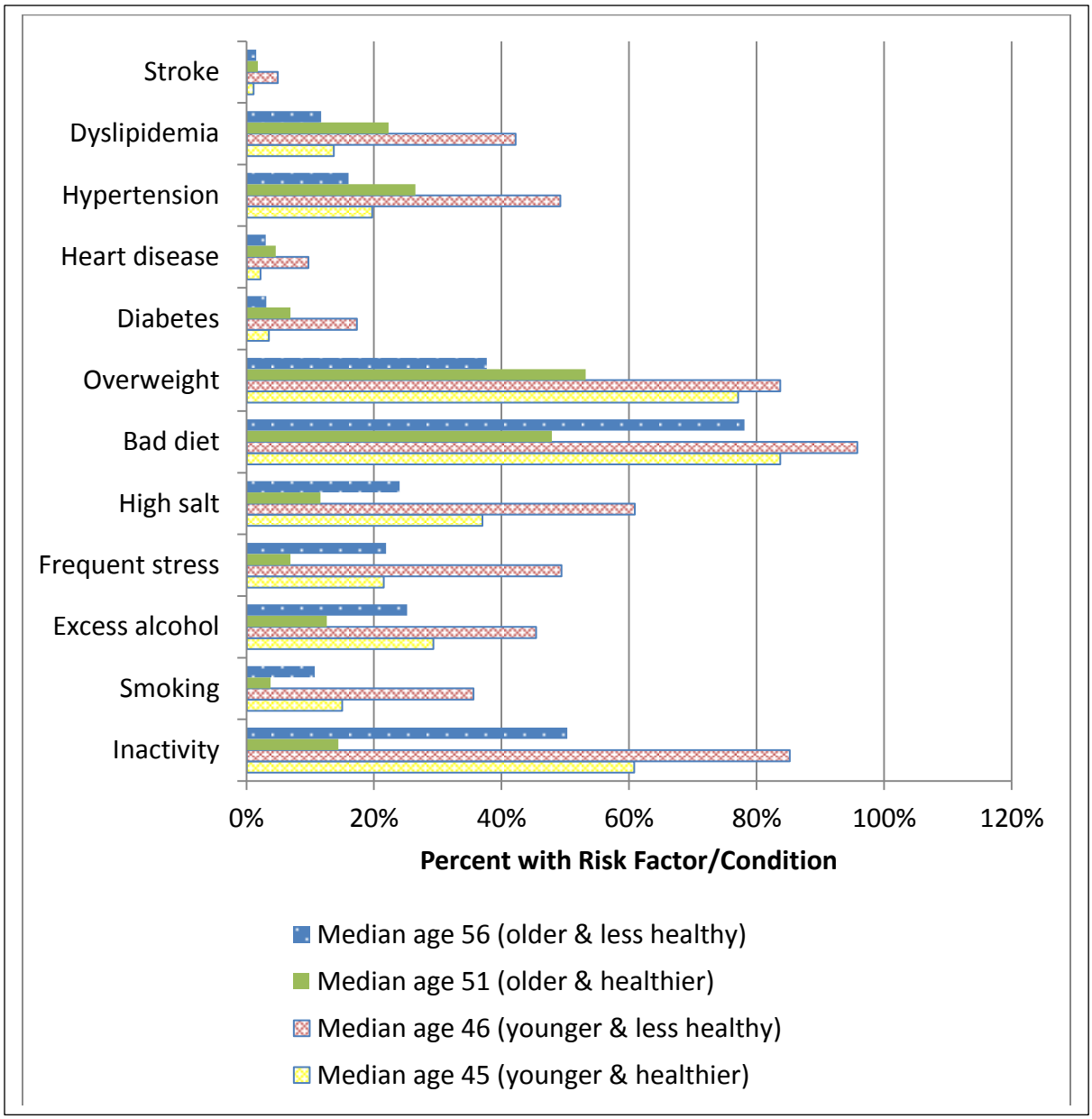
Three- and four-group k-means solutions were reproducible even when the file was split by gender or Air Miles status. A five-group solution could also be generated but was not reproducible when the file was split by gender.

For the four-group k-means solution (K-means Solution 2), clusters sizes were 39.3%, 24.7%, 21.9%, and 14.1%, resulting in a large-to-small ratio of 2.79. All proportions, means and effects sizes for the four groups produced in K-means Solution 2 are provided in Table 5 in Appendix 5.

Groups varied significantly in the distance of cases from cluster centres (ϖ=.257, a large effect). Both clustering variables were associated with large effects, being modestly larger for lifestyle healthiness score (ϖ=.895) than number of health concerns (ϖ=.764). Cluster membership had a large effect for the non-clustering variable of age in years (ϖ = .166); for the categorical variable of age group, effect size approached but did not meet the criteria of a medium-sized effect (Cramer's V_{3df}= .144, whereas a medium-sized effect is defined as $\geq$.170 and < .290).

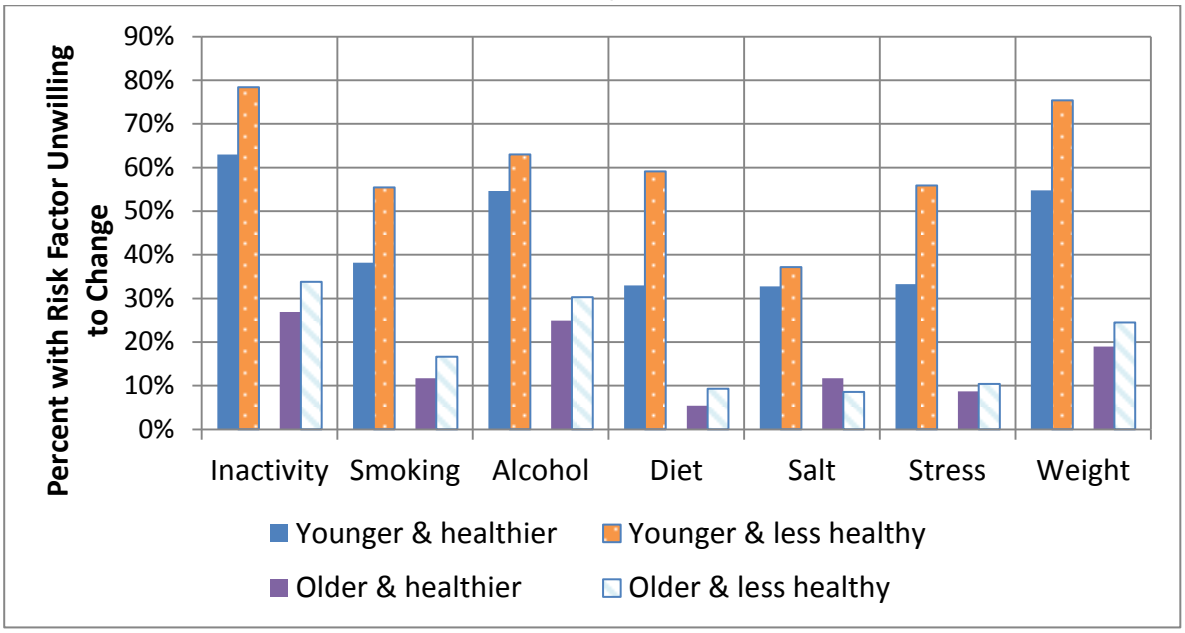
The four groups fell into two broad age ranges: two younger (median ages 45 and 46 years) and two older (median ages 51 and 56 years). Younger groups had higher proportions of modifiable risk factors and lower proportions of vascular conditions than the older groups but within each age dyad there were differences that suggested "healthier" and "less healthy" groups (Figure 27). Group membership had a large effect on physical inactivity (Cramer's V_{1df}=.518) and moderate effects on the report of smoking, bad dietary behaviours, salt intake, stress and obesity (Cramer's V_{1df} >.300 and <.500). The effect on excess alcohol consumption was small (Cramer's V_{1df}=.259).

Figure 27: Proportion by K-means Solution 2 group with modifiable risk factor and vascular condition



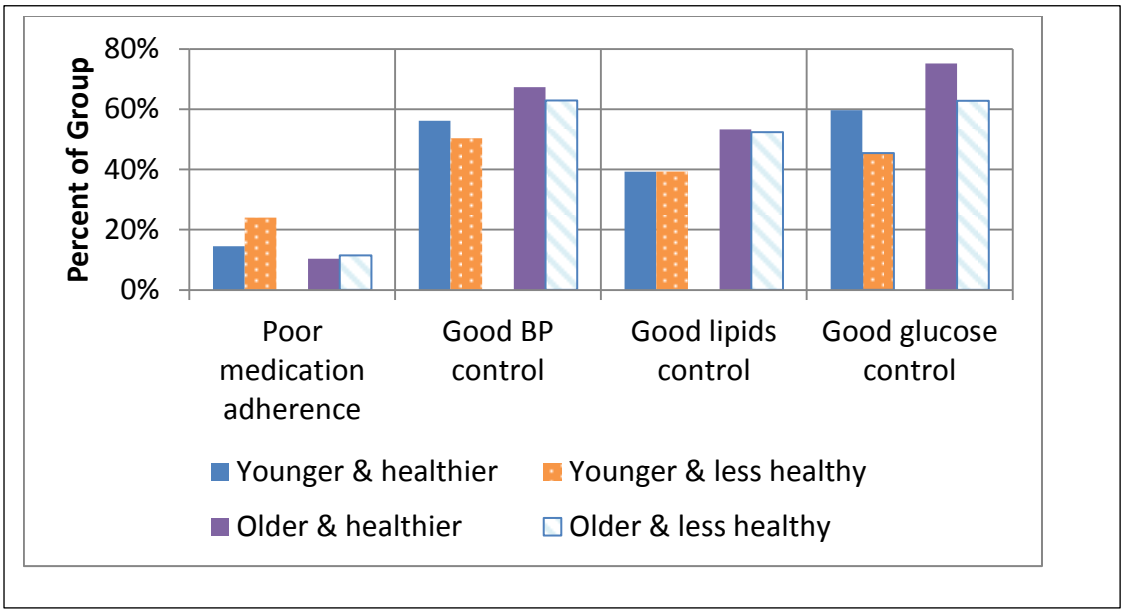
Since the clustering variable of lifestyle healthiness score incorporates stage of change, group membership may reflect readiness to change. As shown in Figure 28, younger age groups had larger proportions unwilling to change, which may reflect the relationship between age and health conscientiousness (32). Within each age dyad, the "less healthy" group had larger proportions unwilling to change. Group membership had a medium-sized effect (Cramer's $V_{1df} \geq .300$ but $<.500$) for all variables except salt consumption which fell just below the cut-off (Cramer's $V_{1df} = .276$; see Table 5 in Appendix 5 for more information).

Figure 28: Proportion unwilling to change by K-means Solution 2 group



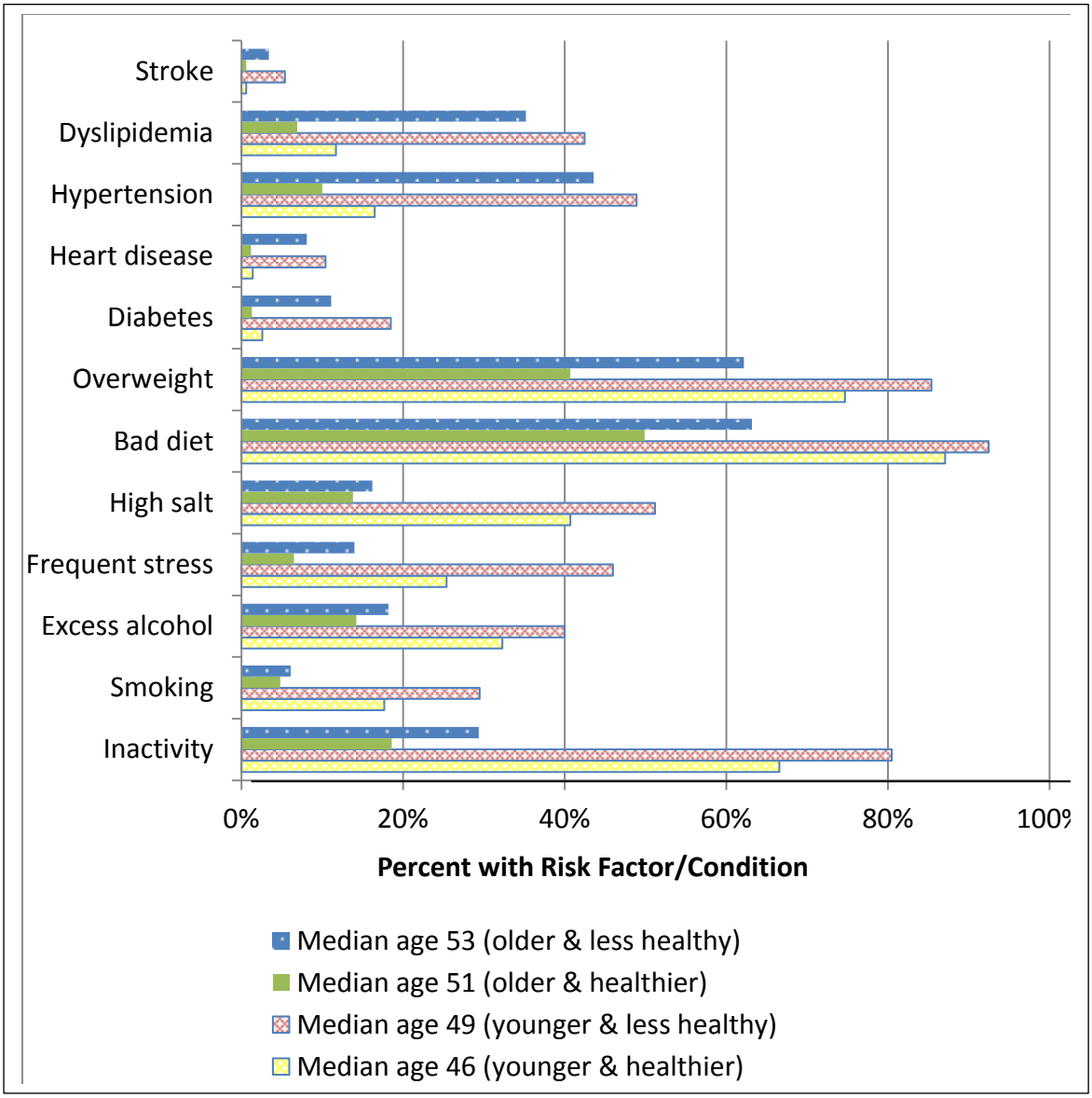
Would differences between clusters be evident in health behaviours not used for clustering? Figure 29 shows self-reported poor medication adherence and good blood pressure, lipids and glucose control for those with diagnosed conditions. With the exception of lipid control, for which there is little or no variation between groups, trends supported the concept of two age dyads that differ in their health consciousness or conscientiousness. However, effect sizes were small (Cramer's V_{1df} were .102 for medication adherence, .078 for blood pressure control, and .121 for glucose control), suggesting only weak relationships with group membership.

Figure 29: Health behaviours by K-means Solution 2 Group



When the two-step procedure was used with these variables, a two-group solution with good cohesion and separation (silhouette co-efficient = 0.6) was generated but a four-group solution (Two-step Solution 2) could be forced that was only slightly weaker (silhouette co-efficient = 0.5). Group sizes were 45.7%, 25.0%, 23.2% and 17.1%, resulting in a large-to-small ratio of 2.03. The four-group solution had good internal consistency, as the silhouette co-efficient remained at 0.5 even when the file was split by gender or Air Miles status. Groups varied by age but in a more linear fashion than in K-means Solution 2, with median ages of 46, 49, 51 and 53 years. Proportions, means and group membership effects sizes for Two-Step Solution 2 are provided in Table 5 in Appendix 5; as well, key variables are summarized in Table 9. Figure 30 shows proportions by group for vascular diseases and modifiable risk factors.

Figure 30: Proportions by Two-Step Solution 2 for modifiable risk factors and vascular conditions



Of variables not used for clustering, there was a large effect for age (ϖ=.144) and a close to medium-sized effect for the age-confounded variable of medication use (Cramer's V_{1df}=.270). Proportions (see Figure 29) and means for vascular disease and modifiable risk factors (Table 9) suggested a similar pattern as seen in K-means Solution 2.

Table 9: Comparison of K-means Solution 2 and Two-step Solution 2

	Group A	Group B	Group C	Group D
K-means Solution 2				
Mean age (years)	44.8	45.2	49.5	53.3
Median age (years)	45	46	51	56
Mean number vascular diseases	0.2	0.8	0.4	1.4
Mean number modifiable risk factors	3.2	4.6	1.4	2.8
Mean number non-modifiable risk factors	1.4	2.5	1.6	3.3
Number of health concerns	4.8	7.9	3.4	7.5
Lifestyle healthiness score	27.0	22.1	32.4	29.3
Proportion of population (%)	24.7%	14.1%	39.3%	21.9%
Two-step Solution 2				
Mean age (years)	45.7	47.5	49.3	51.3
Median age (years)	46	49	51	53
Mean number vascular diseases	0.3	1.3	0.2	1.3
Mean number modifiable risk factors	3.4	4.3	1.5	2.1
Mean number non-modifiable risk factors	1.6	3.3	1.1	3.1
Number of health concerns	5.4	8.9	2.8	6.2
Lifestyle healthiness score	26.1	24.2	31.8	31.1
Proportion of population (%)	25.0%	17.1%	34.7%	23.2%

Although the pattern created by Two-step Solution 2 was similar to that of K-means Solution 2, it was not identical and in some respects less robust. First, lifestyle healthiness score, which incorporates the prevalence of modifiable risk factors and readiness to change, did not vary as much between groups as they did in the k-means solution (effect size for this variable was ϖ=.807 for Two-step Solution 2 compared to .895 for K-means Solution 2). Second, differences between younger groups in their readiness to change modifiable risk factors (see Figure 31) were smaller than in K-means Solution 2, with three not meeting the .30 cut-off suggesting a medium-sized effect (Cramer's V_{1df} were .197 for dietary salt, .241 for smoking and .251 for alcohol; see Table 6 in Appendix 5).

Figure 31: Percent unwilling to change modifiable risk factor by Two-Step Solution 2 group

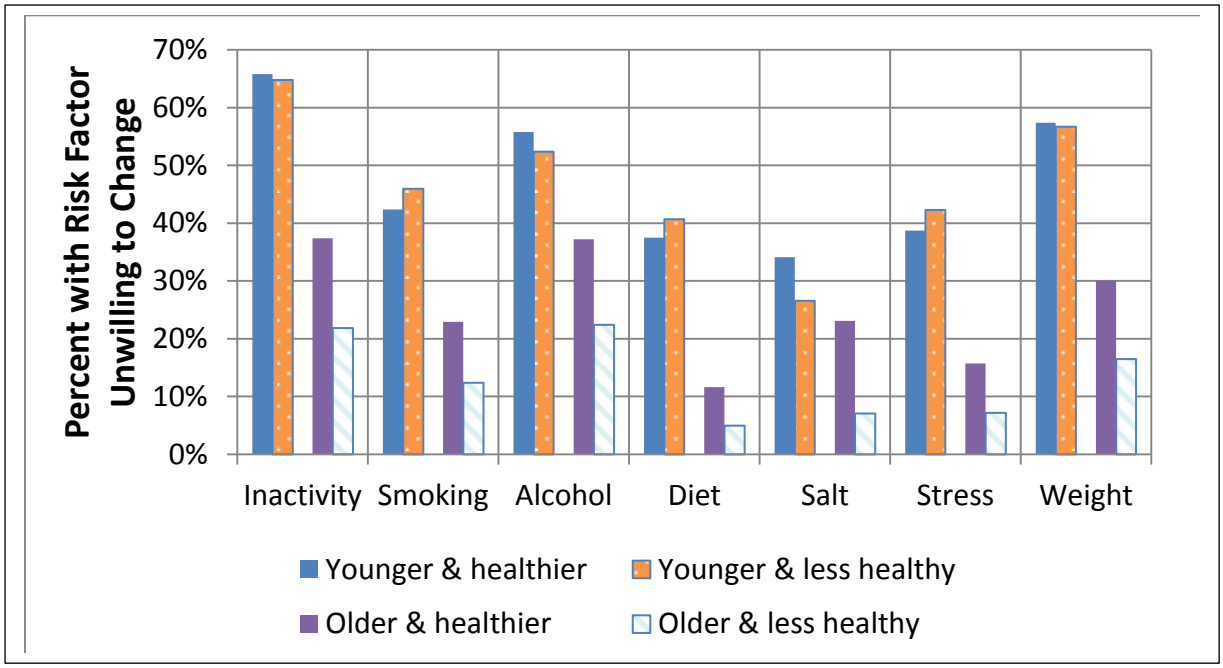


Figure 32 shows the proportions reporting poor medication adherence or good blood pressure, lipids and glucose control. Although the overall pattern was similar to that obtained by K-means Solution 2, differences between groups and thus effect size were smaller. These findings suggest the two-step solution may not be as effective as K-means Solution 2 in developing distinctive sub-groups within the HRA population.

Figure 32: Health behaviours by Two-Step Solution 2 group

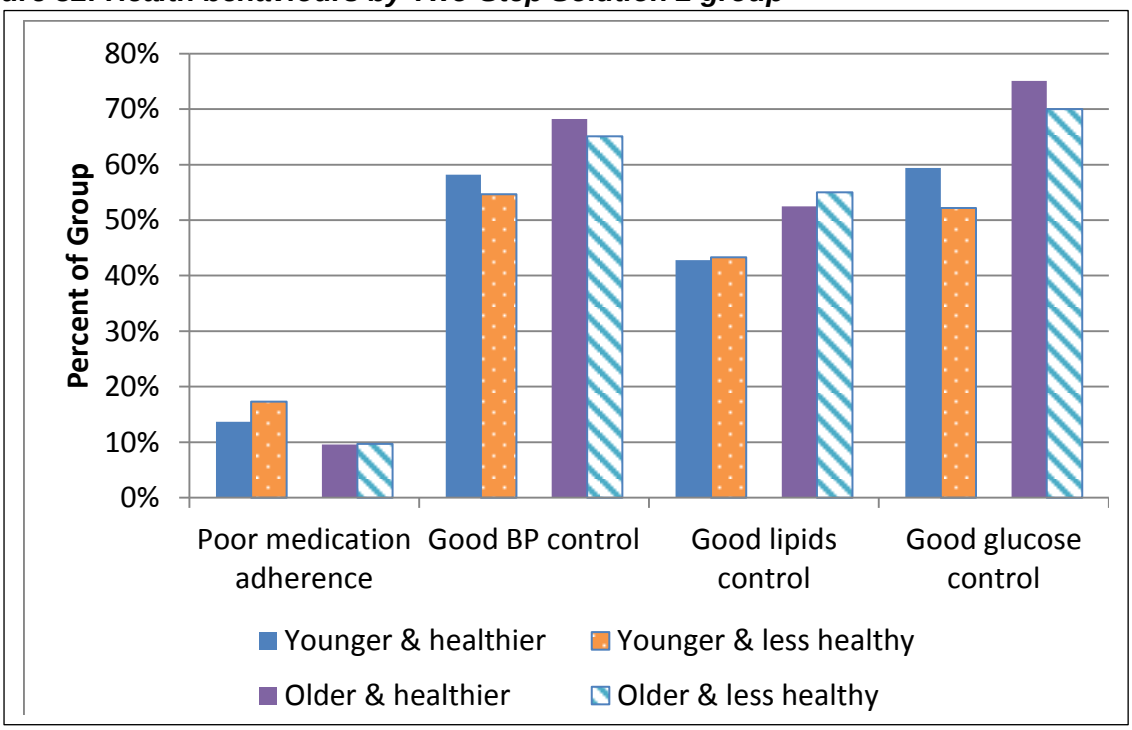


Table 10 summarizes the three segmentations created using the two clustering variables of lifestyle healthiness score and number of health concerns. Both the k-means and two-step procedures could produce four-group solutions with good reproducibility and face validity. Moreover, both produced similar patterns, in which it appears the HRA population consists of groups that vary by age, health status and readiness to make lifestyle changes. Agreement between the two solutions was good (84,854/118,941 or 71.3%, with a Cohen's *kappa* of 0.602), suggesting similarities between the two segmentations did not occur by chance. As noted by Dolnicar, a segmentation may be considered "stable" when it can be reproduced using different clustering procedures (341).

Table 10: Comparison of Segmentation Solutions: Lifestyle Healthiness Score and Number of Health Concerns as Clustering Variables

	K-means Clustering (K-means Solution 2)	**Two-Step Clustering (Two-step Solution 2)**	**Latent Class Analysis**
Potential solutions	3-, 4- and 5-group solutions possible	2-group solution recommended; 4-group possible	None
Internal consistency for 4-group solution	Good	Good	N/A
Group sizes of the selected solution	Range between 14% and 39%	Range between 17% and 35%	N/A
Large-to-small group ratio	2.79	2.03	N/A
Face validity of 4-group solution	Good	Good	N/A
Differentiation in 4-group solution	Large effect for age	Large effect for age	N/A
Distance of cases from cluster centre	Large	N/A	N/A
Silhouette Coefficient (2-step only)	N/A	Fair (0.5)	N/A
Classification error rate (LCA only)	N/A	N/A	N/A
Agreement between 4-group solutions	71.3%, Cohen's *kappa* =0.602, *p*<.001		N/A

Approach 3: Clustering using age, lifestyle healthiness, number of vascular diseases and number of non-modifiable risk factors

Approaches 1 and 2 indicate age is an important variable in distinguishing groups within the HRA population. Therefore, age was added as a clustering variable, along with lifestyle healthiness score, number of vascular diseases and number of non-modifiable risk factors. Number of modifiable risk factors was not used for clustering as it is encompassed within the lifestyle healthiness score. The four clustering variables were

able to generate k-means and two-step solutions; however, a significant two-, three-, four- or five-group solution could not be generated using LCA (see Table 6 in Appendix 5).

When k-means clustering was used, a three-group solution lacked internal consistency in that it could not be reproduced when the file was split. There was no convergence for a five-group solution. In contrast, a four-group solution had good internal consistency, in that it could be reproduced when the file was split.

All proportions, means and effect sizes for the groups formed by K-means Solution 3 are provided in Table 7 in Appendix 5. While the distance of cases from cluster centres varied significantly ($p<.001$), the effect was small ($\varpi = .058$), suggesting this was not a robust solution. Effect sizes of the clustering variables suggested the solution was determined largely by age ($\varpi = .953$), followed by number of vascular diseases ($\varpi = .362$), and lifestyle healthiness score ($\varpi = .193$). Although used as a clustering variable, number of non-modifiable risk factors had only a medium-sized effect ($\varpi = .072$).

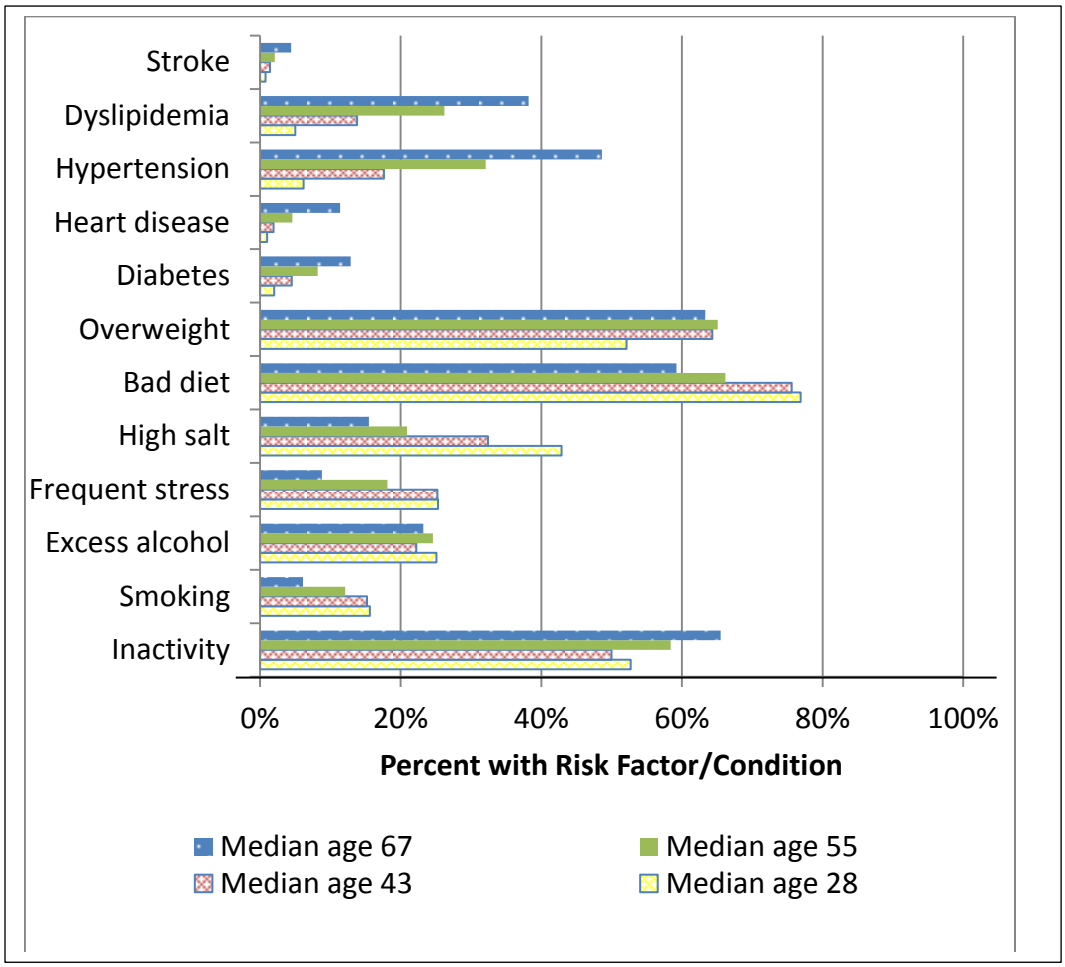
In K-means Solution 3 three variables not used for clustering showed medium effect sizes: medication use (Cramer's $V_{1df} = .297$), working full or part-time (Cramer's $V_{1df} = .420$) and marital status as a binary variable (married/common-law *vs*. not; Cramer's $V_{1df} = .262$). This was the first segmentation for which employment or marital status had more than small associated effect sizes. However, as discussed in Chapter 4, these variables are confounded by age.

For K-means Solution 3 the median ages of groups were almost linear, being 28, 43, 55 and 67 years. As expected, the mean number of vascular diseases increased by age in a linear fashion (see Table 11). Mean number of modifiable risk factors declined with age but differences between groups were modest (see Table 11). Means for the number of modifiable risk factors were similar for the two youngest age groups (2.9 for both) and then declined to 2.1 for the group with the median age of 55 and 1.9 for the group with the median age of 67.

Figure 33 shows the proportions reporting individual vascular diseases and modifiable risk factors by group. Vascular diseases had the expected positive relationship with the median age of the groups. Of the modifiable risk factors, smoking, a bad diet, salt and stress had the expected negative relationship with age but inactivity, alcohol consumption and weight (overweight/obesity) did not. Effect of group membership was consistently small (Cramer's $V_{1df} < .300$), with the exception of hypertension for which there was a medium-sized effect (Cramer's $V_{1df} = .330$). Group membership also had only small effects on readiness to make lifestyle changes or health behaviours such as

medication adherence or blood pressure, lipids or glucose control (for all variables, Cramer's V_{1df} <0.156).

Figure 33: Proportions by K-means Solution 3 group for modifiable risk factors and vascular conditions



When the same clustering variables were used with the two-step procedure, a five-group solution was recommended and had a silhouette co-efficient of 0.4, indicating fair cohesion and separation. Three- and four-group solutions could be forced without changing the silhouette co-efficient; moreover, the four-group solution did not lose cohesion or separation when the file was split by gender or Air Miles status.

For the four-group solution (Two-step Solution 3), groups were fairly equitable in size: 29.6%, 27.7%, 27.5%, and 15.2%, for a large-to-small ratio of 1.95. Means, proportions and effect sizes for all variables by Two-step Solution 3 group are provided in Table 8 in Appendix 5. Effect sizes for the clustering variables were quite different from those observed in K-Means Solution 3: in Two-step Solution 3, the largest effect was associated with number of vascular diseases (ϖ = .698), followed by lifestyle healthiness score (ϖ = .484), number of non-modifiable risk factors (ϖ = .410), and age (ϖ = .223).

These effect sizes suggest that whereas groups in K-means Solution 3 were determined primarily by age, in Two-step Solution 3 they were shaped more by number of vascular diseases. Of variables not used for clustering or closely related to clustering variables, there was only one with even a moderate-sized effect for group membership: being prescribed medication (Cramer's V_{1df} = .388). As previously discussed, this variable may be confounded by age.

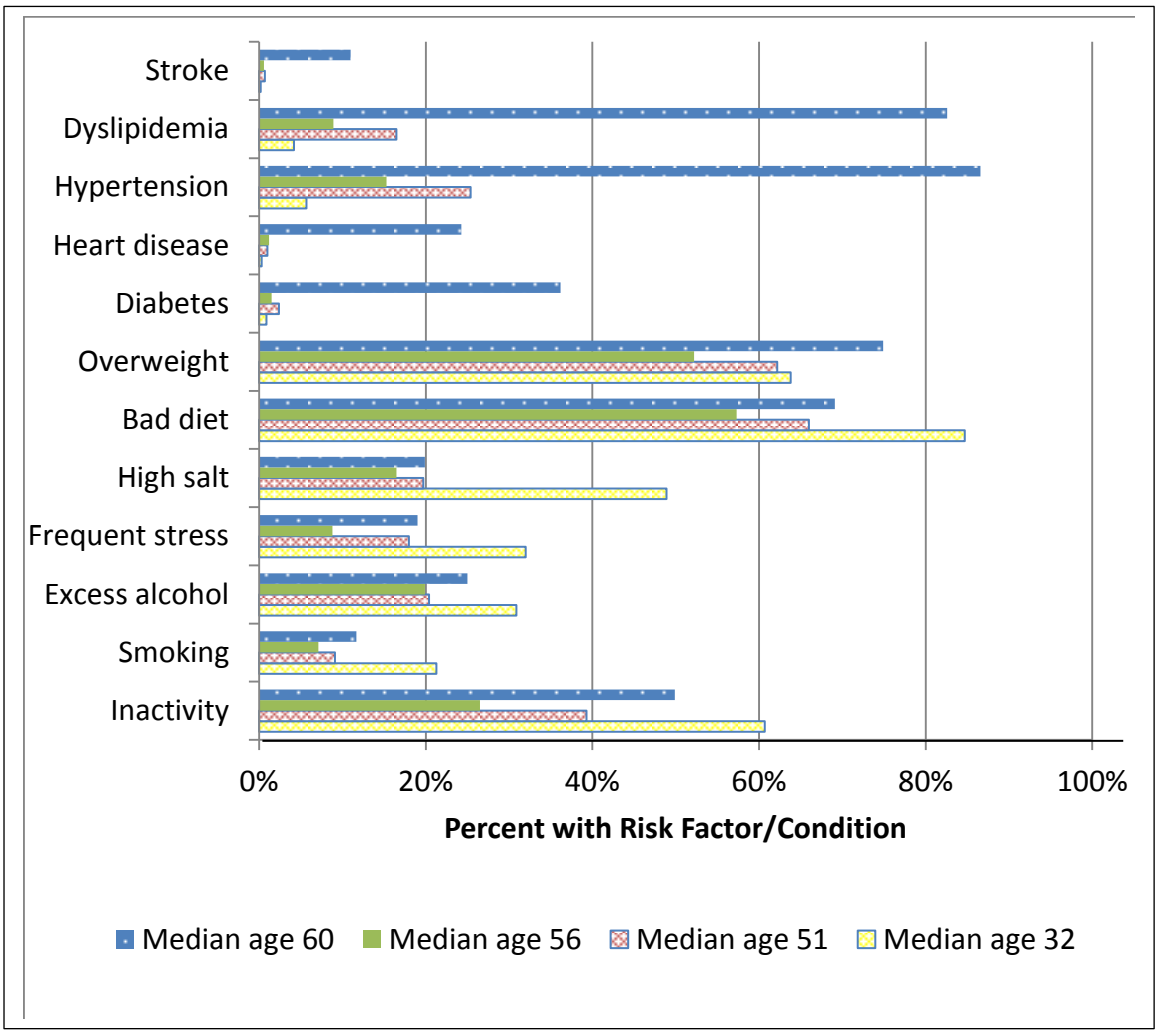
Table 11: Comparison of K-means Solution 3 and Two-step Solution 3

	Group A	Group B	Group C	Group D
K-means Solution 3				
Mean age (years)	28.0	42.7	54.9	68.0
Median age (years)	28	43	55	67
Mean number vascular disease	0.2	0.4	0.7	1.2
Mean number modifiable risk factors	2.9	2.9	2.5	2.1
Mean number non-modifiable risk factors	2.1	2.2	2.1	1.9
Number of health concerns	5.1	5.4	5.3	5.1
Lifestyle healthiness score	28.2	28.2	29.3	30.1
Proportion of population	20.7%	26.5%	34.0%	18.8%
Two-step Solution 3				
Mean age (years)	33.9	50.3	56.3	59.7
Median age (years)	32	51	56	60
Mean number vascular disease	0.1	0.5	0.8	2.5
Mean number modifiable risk factors	3.4	2.3	1.9	2.7
Mean number non-modifiable risk factors	1.9	3.1	0.8	2.8
Number of health concerns	5.5	5.9	2.9	7.9
Lifestyle healthiness score	26.2	29.9	31.0	28.7
Proportion of population	29.6%	27.5%	27.7%	15.2%

In Two-step Solution 3, the mean number of vascular diseases was low for the three younger age groups and then increased substantively for the oldest age group (Table 11). Number of modifiable risk factors and vascular conditions did not follow any clear linear trends (Figure 34), suggesting this solution had poor face validity. Group membership had only small effects on medication adherence or chronic disease management (for all variables, Cramer's $V_{1df} \leq 0.130$). Even though lifestyle healthiness score incorporates readiness to change, the effect of group membership on readiness to change modifiable risk factors was only small-to-moderate in size (Cramer's V_{1df} ranged from a low of 0.192 for inactivity to a high of 0.276 for stress).

When renumbered in order by age, agreement between the two solutions (K-means Solution 3 and Two-step Solution 3) was poor (56,068/110,086 = 47.1%, Cohen's *kappa* = 0.288, $p<.001$). In other words, even when the same variables were used for clustering, the groups formed were significantly different.

Figure 34: Proportions by Two-step Solution 3 for modifiable risk factors and vascular conditions



As shown in Table 12, though the k-means and two-step clustering solutions had good internal consistency, in each case face validity was questionable. Moreover, neither solution appeared to offer new or additional insights into the HRA population above and beyond that available through analysis by age.

Table 12: Comparison of Segmentation Solutions: Age, Lifestyle Healthiness Score and Number of Vascular Diseases and Non-modifiable Risk Factors as Clustering Variables

	K-means Clustering (K-means Solution 3)	Two-Step Clustering (Two-step Solution 3)	Latent Class Analysis
Potential solutions	3-group solution lacked internal consistency; no 5-group solution possible	2-group solution recommended; 4-group possible	None
Internal consistency of 4-group solution	Good	Good	N/A
Group sizes of the selected solution	Range between 19% and 34%	Range between 17% and 35%	N/A
Large-to-small group ratio	1.80	1.95	N/A
Face validity	Good	Poor	N/A
Differentiation	Medium-sized effect for medication, working full/part-time and marital status	Medium-sized effect for medication use	N/A
Distance of cases from cluster centre	Small	N/A	N/A
Silhouette Coefficient (2-step only)	N/A	Fair (0.4)	N/A
Classification error rate (LCA only)	N/A	N/A	N/A
Agreement between 4-group solutions	47.1%, Cohen's *kappa* =0.288, *p*<.001		N/A

Approach 4: Modifiable and non-modifiable risk factors as nominal variables

In most of the previous segments, number of modifiable and non-modifiable risk factors appeared to play important roles in group formation. What effect would different individual risk factors have on the formation of groups? It is possible, for example, that certain modifiable risk factors may cluster together, thereby forming groups of different types of HRA users (e.g., those who both smoke and drink alcohol to excess may form one group, while those who report poor dietary behaviours may form another). Nominal variables cannot be analyzed using k-means clustering but can be accommodated in LCA and two-step clustering. Approach 4 was therefore initiated by seeing which modifiable and non-modifiable risk factors could form statistically significant LCA solutions.

For the first attempt, as described in Chapter 4, the seven modifiable risk factors (physical inactivity, ≥1 poor dietary behaviour, frequent stress, overweight/obesity,

smoking, and excessive salt or alcohol consumption) were coded to incorporate stage of change (e.g., from 1 for Precontemplation to 5 for Maintenance). These seven variables were then entered as ordinal clustering variables for LCA. For a four-group solution, bivariate residuals (BVRs) were large, with the largest being for interaction between weight and inactivity (256.9), diet and salt (131.9), and smoking and excess alcohol consumption (118.6). Variance left unexplained was also large (L^2=50,145.8). Individual variables were eliminated according to their BVRs and variance explained in an attempt to generate a statistically significant solution. Although numerous attempts were made, even when controlling for interactions between variables, no combination of variables was found that could produce a statistically significant solution.

A different approach was thus used, in which modifiable and non-modifiable risk factors were used for LCA as nominal but binary variables (present/absent). Based on amount of variance explained and BVRs, variables were individually eliminated until statistically significant segmentations were generated (Table 13). This process identified five variables: three were diet-related (fruit and vegetable, fish and salt consumption) and two concerned family medical history (family history of premature heart disease and of high cholesterol).

Table 13: Latent class analysis using modifiable and non-modifiable risk factors

Number of Clusters	***LL***	***BIC(LL)***	***Number Para-meters***	***L²***	***Degrees of Freedom***	***p-value***	***Classifica-tion Error***
2	-387707.93	775544.448	11	5343.382	20	9.6e-1136	0.1709
3	-385490.91	771180.536	17	909.335	14	4.3e-185	0.2179
4	-385043.40	770355.651	23	14.315	8	0.074	0.2974
5	-385038.03	770415.048	29	3.578	2	0.17	0.3610

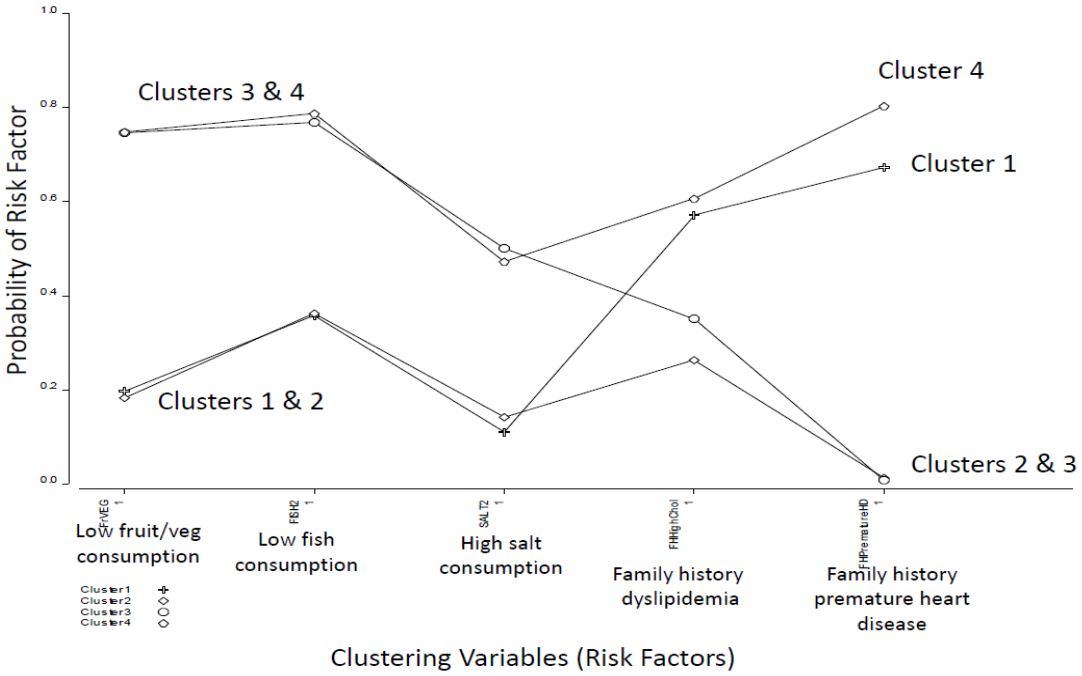
As shown in Table 13, both four- and five-group solutions were statistically significant (in Latent Gold, a statistically significant model is indicated by $p \geq 0.05$). The L^2 (a measure of the amount of association between variables that remains unexplained) was smaller for the five-group solution but the four-group solution could be considered a better fit because it has fewer parameters (23 *vs*. 29) and a smaller classification error rate (29.7% *vs.* 36.1%). Bootstrapping found the five-group solution was not a significant improvement over the four-group model (p=0.16, SE=0.02) and the p value for the four-group model was probably under-estimated (p= 0.215, SE=0.02).

Figure 35 illustrates the conditional probability of group membership according to the five clustering variables; values are also provided in Table 9 in Appendix 5. Figure 33 shows that for dietary risk factors, the two younger group, Clusters 3 (median age 45

years) and 4 (median age 47), had higher probabilities of unhealthy lifestyle choices than Clusters 1 and 2 (both of which had median ages of 52). The two younger groups varied, however, when it came to the two non-modifiable risk factors. Cluster 3 had a moderate probability of a family history of dyslipidemia and a low probability of a family history of premature heart disease (370). Cluster 4, on the other hand, had higher probabilities of both non-modifiable risk factors.

The two older clusters, Clusters 1 and 2, had low probabilities of poor dietary behaviour. However, as with the two younger groups, they diverged in relationship to the non-modifiable risk factors: Cluster 1 had high probabilities of the two non-modifiable risk factors whereas Cluster 2 had low probabilities.

Figure 35: Probabilities of Risk Factors Being Present by LCA Solution 1 Cluster



Proportions, means and effect sizes for group membership when the probabilities of group membership were applied to the HRA sample are reported in Table 10 in Appendix 5. As the LCA procedure placed all persons with a family history of heart disease in either Cluster 4 or 1, it had a large associated effect size (Cramer's V_{1df}=.956), even though only half (47.6%) of all HRA users reported this risk factor. Effect sizes associated with the other clustering variables were large (Cramer's V_{1df} were .680 for fruit and vegetable consumption, .566 for fish consumption and .554 for salt consumption), with the exception of family history of dyslipidemia, which had a small-to-medium-sized effect (Cramer's V_{1df}=.261). Of variables not used for clustering, there was a large-sized effect for only one: age in years (ϖ=.612). Effects sizes for other variables were small.

Figure 36: Proportions by LCA Solution 1 groups for modifiable risk factors and vascular conditions

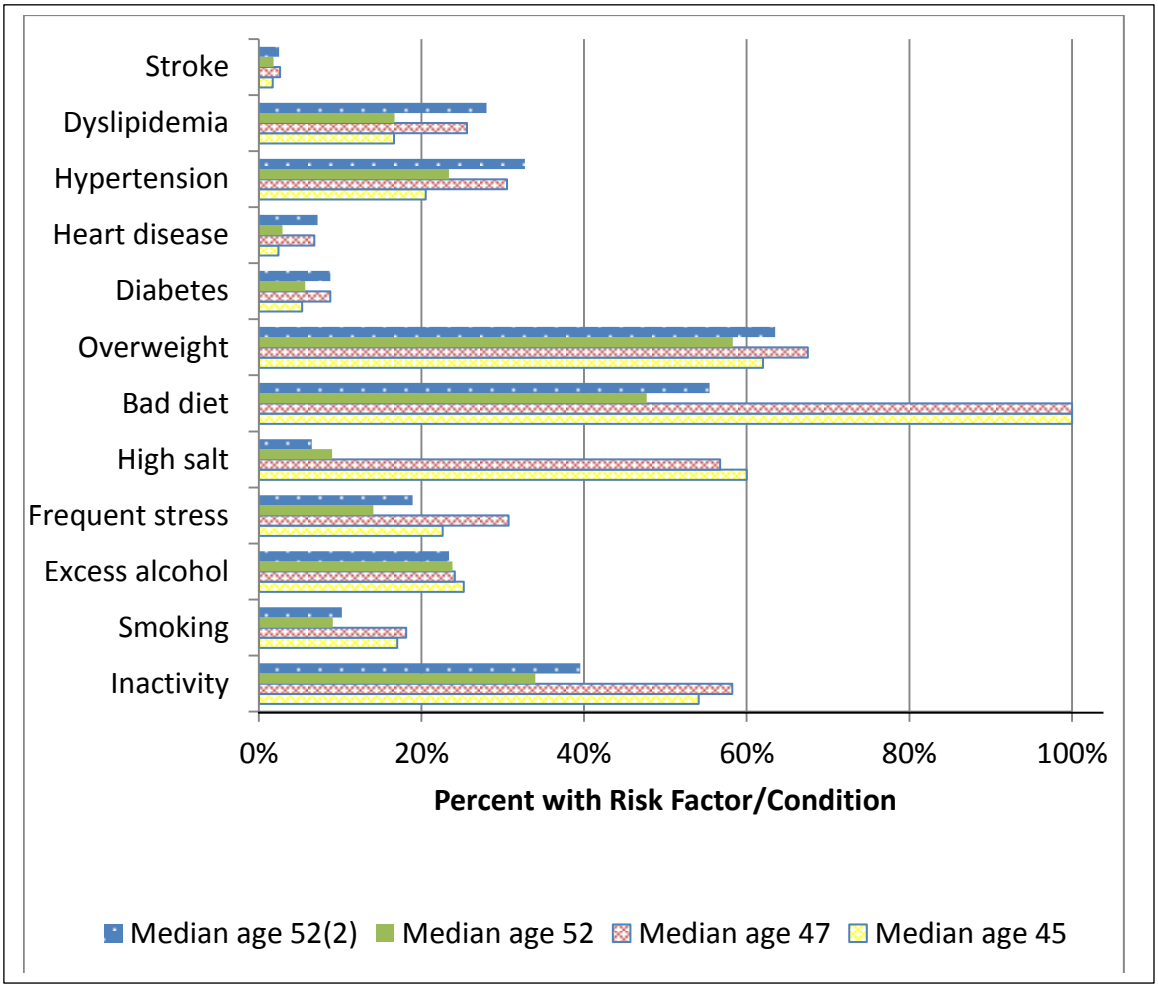


Figure 36 shows proportions by LCA Solution 1 for modifiable risk factors and vascular conditions. If the groups were divided into two broad age dyads (two with a median age of 52 and two with lower median ages of 45 and 47 years), one of the pairs consistently had higher rates of vascular diseases than the others. Could the pattern of "more healthy" compared to "less healthy" groups seen in K-means Solution 2 and Two-step Solution 2 be present? The answer appears to be no, as there were inconsistencies in the patterns for modifiable risk factors. For example, the youngest group (median age 45) had lower proportions than the group with the median age 47 in the proportions reporting inactivity, smoking, frequent stress and overweight/obesity, but either no difference or a slightly larger proportions reporting excess alcohol consumption and bad diet. Differences between the two older age groups were also small and inconsistent. In addition, effect sizes for group membership on unwillingness to change modifiable risk factors were uniformly small (Cramer's $V_{1df} \leq .213$).

The amount of variance explained by this model was modest and not always significant; the highest variance explained was for family history of premature heart

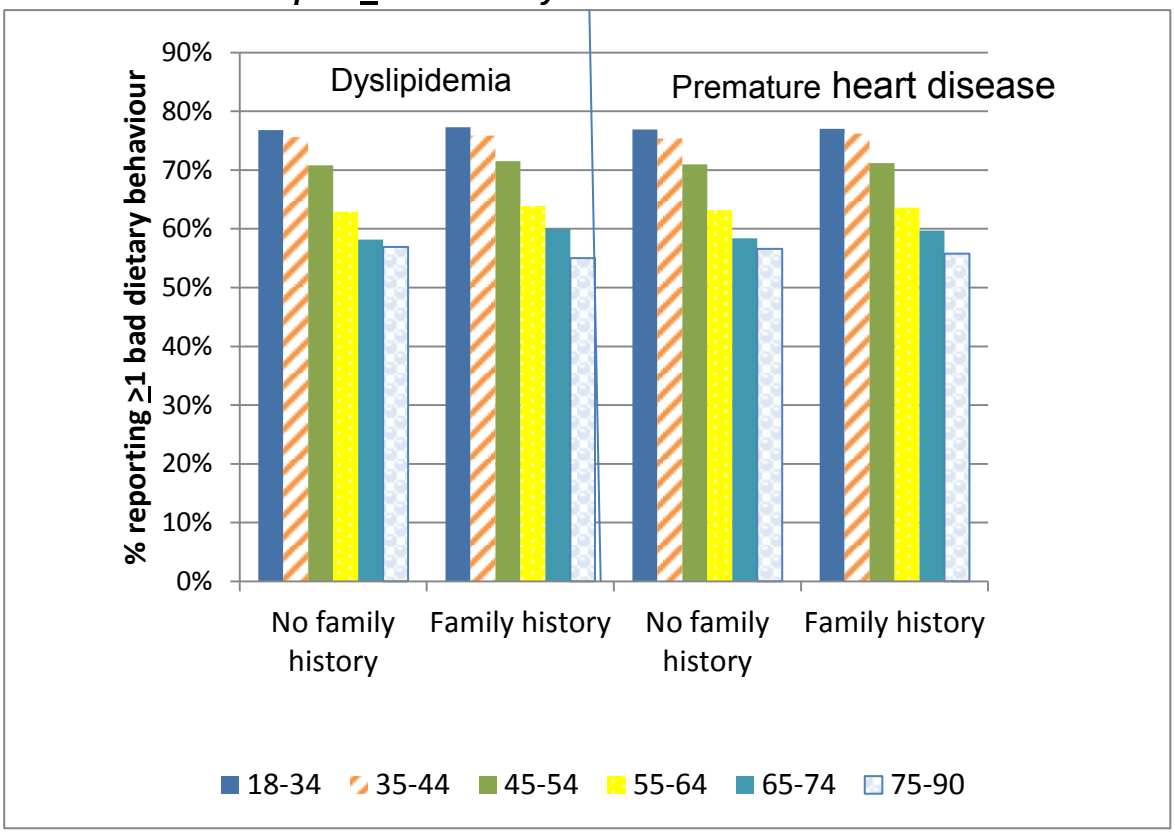
disease (54.7%, *p*=0.806). Although there was a large-sized effect for the non-clustering variable of age in years (ϖ =.612), there was little variance in mean or median ages of groups. Internal consistency was problematic; when the file was split by gender or Air Miles status convergence could not be achieved for both sub-sets. This suggests the four-group LCA solution did not apply equally well to various parts of the HRA population.

The four-group solution had a relatively large classification error rate of 29.7%. The error classification rate reflects the fact that cluster or group membership is not a binary state (yes/no) but an estimation of the most likely class based upon a) the estimated prevalence of each group and b) the *a posteriori* probabilities of a case's membership in each class in comparison to the pattern of responses (i.e., variables) of the group. Cases have probabilities of belonging to more than one group but are assigned to the group to which they have the highest *a posteriori* probability (371). When group membership probabilities were applied to the HRA population, group sizes varied from those in the model (e.g. the size of Cluster 1 changed from 33.2% to 32.2% of cases).

Did the groups form because a family history of dyslipidemia or premature heart disease confounded dietary behaviours? Seventy percent (70.4%) of those with a family history of dyslipidemia reported $\geq$1 bad dietary behaviour compared to 68.9% of those without a history (Cramer's V_{1df}=.016). Likewise, of those with a family history of premature heart disease, 69.7% report bad dietary behaviours compared to 69.5% of those without a family history (Cramer's V_{1df} =.002). In fact, as shown in Figure 37, dietary behaviour was more strongly influenced by age than by family medical history. Regardless of family history, the prevalence of bad dietary behaviours decreased with age; this trend mirrors the large inverse relationship between age and poor dietary behaviours reported in Chapter 4 (*eta*=.139).

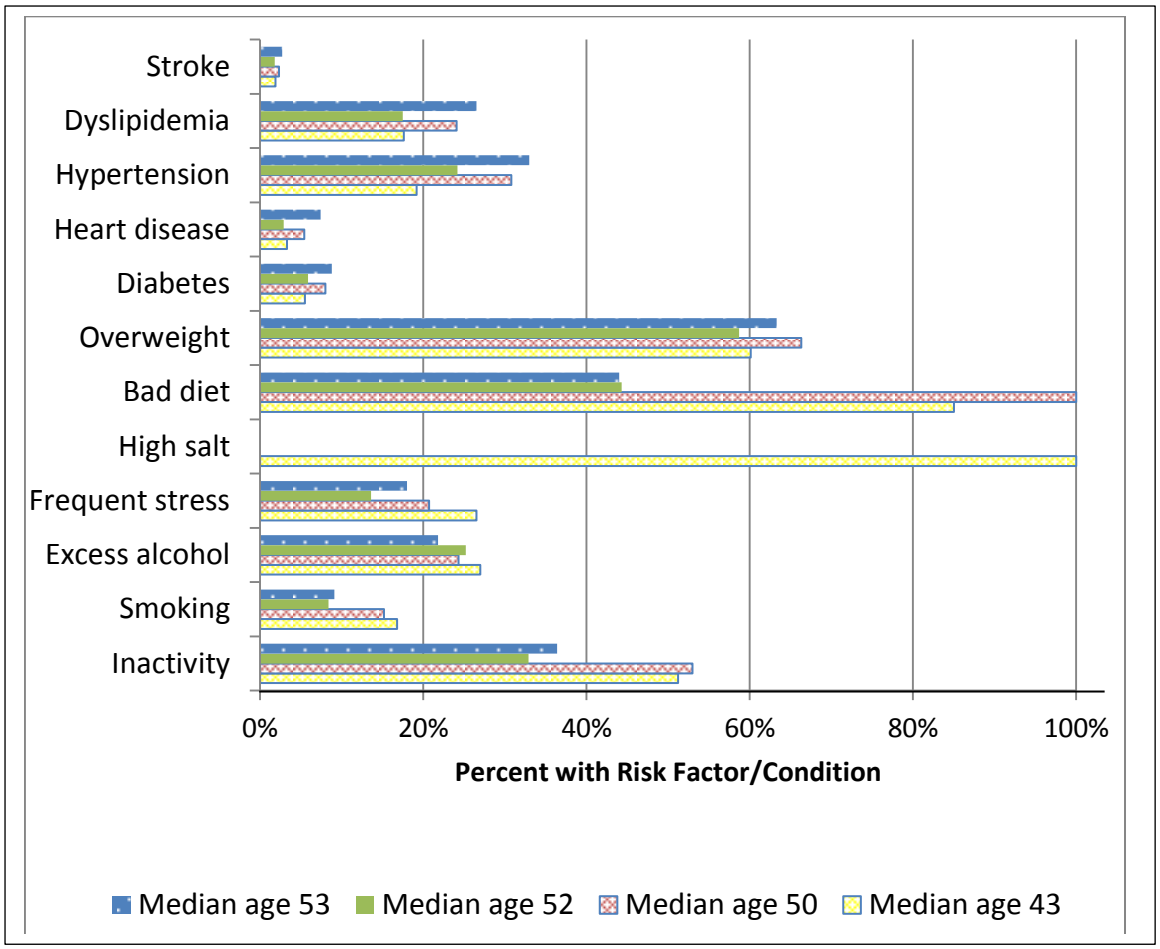
When the five categorical variables identified by LCA Solution 1 were submitted to two-step clustering and the number of groups left unspecified, a four-group solution was generated with fair cohesion and separation (silhouette coefficient = 0.4). A four-group solution had good internal consistency, as when the database was split by gender or Air Miles status there was no change to the silhouette co-efficient. Sizes of the groups formed were fairly equitable, with a large-to-small group ratio of 1.63.

Figure 37: Proportion of people with a family history of dyslipidemia or premature heart disease who report ≥1 bad dietary behaviour



Proportions, means and effect sizes for Two-step Solution 4 are provided in Table 11 in Appendix 5. The two-step procedure put all users reporting high salt consumption in one group (Cluster 2), low fruit and vegetable consumption in either Cluster 2 or 4, and those reporting a family history of heart disease in either Cluster 2, 4 or 1. As a result, salt consumption correlated perfectly with group membership (Cramer's V_{1df}=1.000), while there were large effects for fruit and vegetable consumption (Cramer's V_{1df}=.854) and family history of heart disease (Cramer's V_{1df}=.686). Low fish consumption (Cramer's V_{1df}=.239) and family history of dyslipidemia (Cramer's V_{1df}=.141) had small effect sizes. The formation of groups in Two-step Solution 4 was thus more influenced by salt and fruit and vegetable consumption than in LCA Solution 1, where the primary influence was family history of heart disease.

Figure 38: Proportions by Two-step Solution 4 groups for modifiable risk factors and vascular conditions



The only non-clustering variable to have even a medium-sized effect was age (ϖ=.231). Median ages of the four groups in Two-step Solution 4 were 43, 50, 52 and 53 years. As shown in Table 14, differences in the mean number of vascular diseases between groups were small, although the effect of group membership was medium-sized (ϖ =.129). Mean number of modifiable risk factors was highest for the youngest age group and declined by age group in a non-linear fashion, although effect of group membership was large (ϖ=.555).

As with LCA Solution 1, it was difficult to understand how the groups of similar age differed from one another. For example, as shown in Figure 38, although it appeared that two groups (those with the median ages 53 and 50) had the highest proportions reporting vascular diseases, there was no consistent pattern for modifiable risk factors. In addition, as shown in Table 12 in Appendix 5, the effect of group membership on readiness to change modifiable risk factors was consistently small (Cramer's $V_{1df} \leq .232$).

Table 14: Comparison of LCA Solution 1 and Two-step Solution 4

	Group A	Group B	Group C	Group D
LCA Solution 1				
Mean age (years)	44.6	46.2	50.5	51.0
Median age (years)	42	47	52	52
Mean number vascular disease	0.5	0.8	0.5	0.8
Mean number modifiable risk factors	3.4	3.6	2.0	2.2
Mean number non-modifiable risk	1.4	3.3	3.1	1.3
factors	5.3	7.5	3.8	6.1
Number of health concerns	26.8	26.6	30.0	30.6
Lifestyle healthiness score	24.1%	14.4%	36.1%	25.4%
Proportion of population				
Two-step Solution 4				
Mean age (years)	43.4	49.2	51.0	51.5
Median age (years)	43	50	52	53
Mean number vascular disease	0.5	0.7	0.5	0.8
Mean number modifiable risk factors	3.7	2.8	1.8	1.9
Mean number non-modifiable risk	2.0	2.8	1.4	1.9
factors	6.2	5.6	3.7	5.8
Number of health concerns	26.3	28.3	30.9	30.6
Lifestyle healthiness score	27.5%	25.7%	29.1%	17.8%
Proportion of population				

Table 15 compares the two solutions. Of the two, LCA Solution 1 had better internal consistency but poorer face validity and, as discussed, had a number of weaknesses; however, Two-step Solution 4 did not appear to be more informative or helpful in understanding the HRA population. When groups were renumbered by progressive age, agreement between LCA Solution 1 and Two-step Solution 4 was modestly above chance (80,496/119,264 = 67.5%, Cohen's *kappa*=.565, *p*<.001).

Table 15: Comparison of Segmentation Solutions: Fruit and Vegetable, Fish and Salt Consumption and Family history of Dyslipidemia or Premature Heart Disease as Clustering Variables

	K-means Clustering	Two-Step Clustering (Two-step Solution 4)	LCA (LCA Solution 1)
Potential solutions	None	5-group solution generated; 4-group acceptable	4 or 5-group solutions are significant
Internal consistency	N/A	Poor	Good
Group sizes of four-group solution	N/A	18% to 29%	Theoretically: 19% to 32%; When applied: 14% to 36%
Large-to-small group ratio	N/A	1.63	Theoretical: 1.67; Applied: 2.51
Face validity of 4-group solution	N/A	Questionable	Fair
Differentiation of 4-group solution	N/A	Large effect for age	Large effect for age
Distance of clusters from centre	N/A	N/A	N/A
Silhouette Coefficient (2-step only)	N/A	Fair (0.4)	N/A
Classification error rate (LCA only)	N/A	N/A	29.7%
For 4-group solution, agreement between procedures	N/A	65.7%, Cohen's *kappa* =0.565, $p<.001$	

Approach 5: Modifiable risk factors and vascular diseases as nominal binary variables.

Approach 5 was similar to Approach 4 in that nominal binary variables were entered and then progressively eliminated based on their ability to explain variance, BVR values, and the statistical significance of solutions. Whereas Approach 4 utilized non-modifiable and modifiable risk factors, in Approach 5 the clustering variables were modifiable risk factors and vascular diseases.

LCA was initiated using 16 clustering variables: presence or absence of diabetes, hypertension, dyslipidemia, heart disease, stroke, physical inactivity, smoking, frequent stress, frequent fatty food, fast food or salt consumption and infrequent fruit and vegetable and fish consumption, high salt consumption, excessive alcohol consumption, and overweight/obesity. Significant solutions were generated when the variables were

reduced to five: three dietary (high fat and fast food consumption and salt intake) and two vascular diseases (hypertension and dyslipidemia). As shown in Table 16, significant four- and five-group solutions were generated. The four-group solution had a low L^2 value (15.5) and a low classification error rate (13.6%), while the five-group solution had an even lower L^2 value (1.6) but a higher error classification rate (18.8%).

Table 16: Latent Class Analysis output for modifiable risk factors and vascular diseases

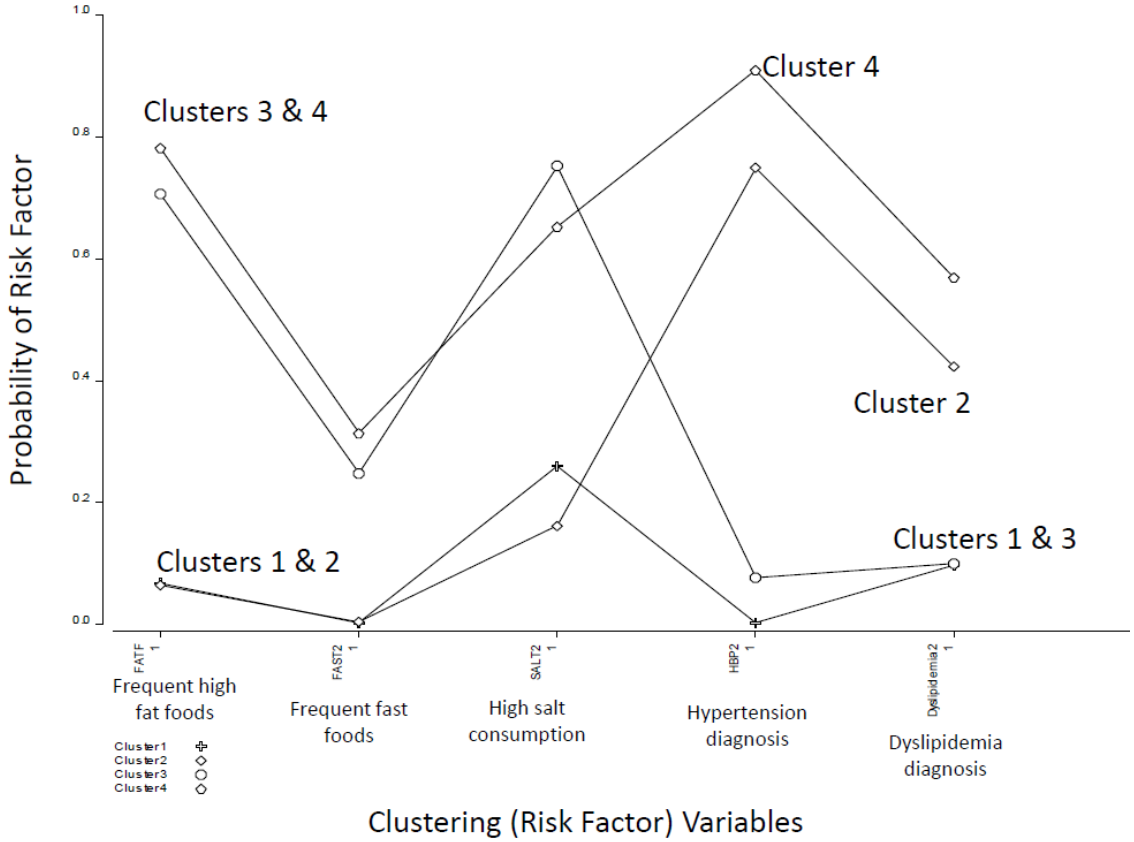
Number of Clusters	LL	BIC(LL)	Npar	L^2	Degrees of Freedom	*p*-value	Classification Error
2	-256627.547	513383.7262	11	10597.2445	20	6.3e-2274	0.1400
3	-251677.441	503553.6782	17	697.0336	14	1.1e-139	0.1420
4	-251336.663	502942.2835	23	15.4760	8	0.051	0.1360
5	-251329.714	502998.5491	29	1.5788	2	0.45	0.1876

LCA Solution 2 had poor internal consistency. When the data base was split by gender, a four-group solution was significant for females (p=0.9) and had a low error classification rate of 13.8% but was not significant for males (p=0.0016), even when it was bootstrapped (bootstrap p = 0.002, SE=0.002; classification error rate=15.8%). Likewise, when the file was split by Air Miles status, a four-group solution was significant for Air Miles participants (p=0.05, error classification rate = 12.5%) but not for non-Air Miles users (p=0.015, error classification rate = 16.3%; bootstrapped p=0.016, SE=0.006).

For the four-group solution (LCA Solution 2), group sizes were 58.1%, 31.9%, 8.4% and 1.6%, giving a large-to-small group ratio of 36.3. Writing tailored messaging for small groups may not be cost-efficient, suggesting that this may not be practical solution for program operators.

Figure 39 illustrates the probabilities of group membership by clustering variables by LCA Solution 2 groups; probabilities are also provided in Table 12 in Appendix 5. Cluster 1, the largest group, consisted of cases in which there were low probabilities of poor dietary behaviour or of hypertension or dyslipidemia. Cases in Cluster 2 had low probabilities of poor dietary behaviour but higher probabilities of hypertension or dyslipidemia. Cluster 3, which accounted for 8.4% of cases, had high probabilities of high fat food and high salt consumption and low probabilities of fast food consumption and of hypertension or dyslipidemia. Cluster 4, which represented only 1.6% of cases, was similar to Cluster 3 in terms of dietary behaviours but also had high probabilities of hypertension and dyslipidemia.

Figure 39: Probability of group membership (LCA Solution 2)



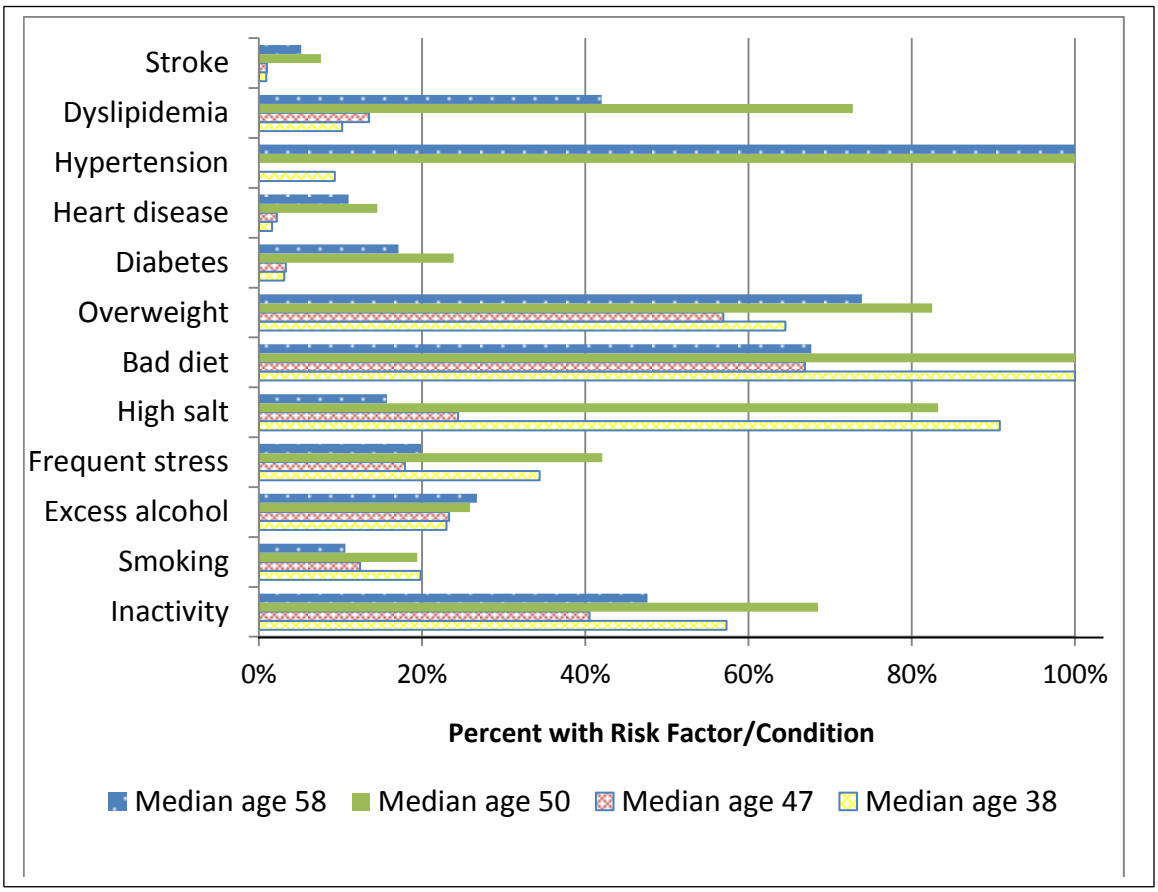
Of the five clustering variables, hypertension had the greatest effect on group membership (Cramer's V_{1df}=.984), as no people with hypertension were placed in the second-youngest group (Cluster 1) and few (9.3%) in the youngest group (Cluster 3) while all persons in Clusters 2 and 4 had hypertension. The second-largest effect (Cramer's V_{1df}=.664) was for fast food consumption: Cluster 1 had no people reporting this dietary behaviour and very few (0.4%) were placed in Cluster 2.

Proportions, means and effect sizes when the probabilities of LCA Solution 2 membership were applied to the HRA data base are provided in Table 13 in Appendix 5. The size of the clusters changed (e.g., Cluster 1 increased from 58.1% to 67.7% while Cluster 2 decreased from 31.9% to 24.5%). Of variables not used for clustering, there was a large effect for age (ϖ =.373), as well as medium effect for medication use (Cramer's V_{1df} = .448) and family history of hypertension (Cramer's V_{1df} =.300).

Figure 40 illustrates proportions reporting vascular diseases and modifiable risk factors by LCA Solution 2 group. There were few consistent trends between groups. For example, the second-oldest group (median age 50) had the largest proportions reporting four of the five vascular diseases (stroke, dyslipidemia, heart disease and diabetes), as well as three of the seven modifiable risk factors (overweight/obesity, frequent stress, and

physical inactivity). Group membership had only a small effect on the report of vascular diseases and modifiable risk factors not used for clustering (for diabetes, Cramer's V_{1df}=.245 but all others were $\leq$.194). Likewise, group membership had only a small effect on readiness to change modifiable risk factors (Cramer's $V_{1df} \leq .151$).

Figure 40: Proportions by LCA Solution 2 groups for modifiable risk factors and vascular conditions

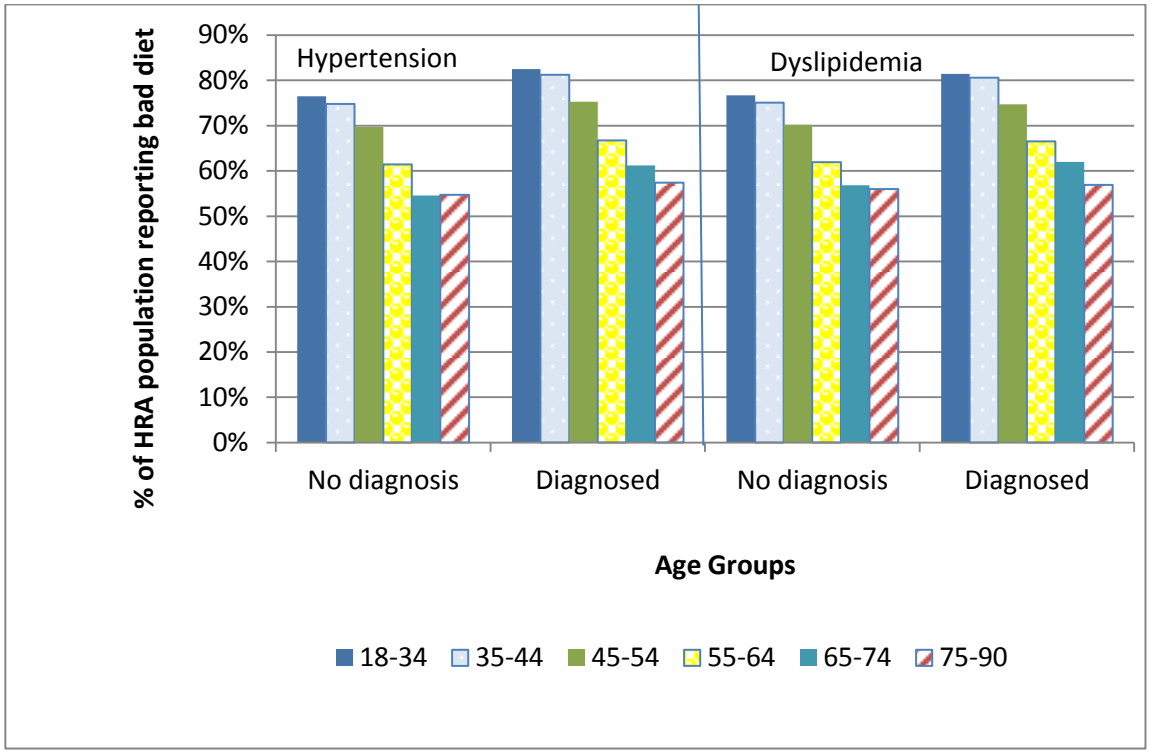


It is possible a diagnosis with hypertension of dyslipidemia confounds dietary behaviour. Table 17 shows there were no consistent or strong relationships between dietary habits and the diagnosis of dyslipidemia or hypertension. Rather, as shown in Figure 41, dietary behaviours were influenced more by age than by chronic disease status. These findings reflect results from a longitudinal Canadian study which found a diagnosis of heart disease, hypertension or diabetes did not increase the proportion of adults aged 50 or older who met the recommended number of servings per day of fruit and vegetables (372). In the previous section, it was proposed that age had a stronger effect on dietary behaviours than family history of dyslipidemia or premature heart disease. Figure 41 suggests a similar relationship exists for dietary behaviours and personal medical history.

Table 17: Prevalence of dietary behaviours by dyslipidemia or hypertension status

Dietary Behaviour	Dyslipidemia			Hypertension		
	No	**Yes**	**Cramer's V (*p*)**	**No**	**Yes**	**Cramer's V (*p*)**
Frequent high fat foods	13.3%	12.1%	0.015 (<.001)	13.5%	11.9%	0.021 (<.001)
High salt	28.6%	23.3%	0.049 (<.001)	30.1%	20.3%	0.096 (<.001)
Frequent fast foods	2.9%	2.6%	0.008 (0.003)	2.9%	2.7%	0.006 (0.037)
Low fruit/vegetable	41.6%	44.1%	0.020 (<.001)	41.7%	43.5%	0.016 (<.001)
Low fish	53.7%	51.3%	0.019 (<.001)	53.6%	52.2%	0.012 (<.001)
≥1 bad dietary behaviour	69.6%	69.3%	0.002 (0.414)	69.6%	69.4%	0.002 (0.561)

Figure 41: Prevalence of ≥1 bad dietary behaviour by hypertension and dyslipidemia status and age group



The utility of LCA Solution 2 for tailoring program messages is questionable. As shown in Table 18, those with the greatest mean number of vascular diseases and modifiable and non-modifiable risk factors fell into a group that comprised less than 1% of the total population. The largest group, which accounted for over two-thirds of cases, had the lowest mean number of vascular diseases and risk factors, making it difficult to determine what sort of health promotion messages would be appropriate.

Table 18: Comparison of LCA solution 2 and two-step solution 5

	Group A	Group B	Group C	Group D
LCA Solution 2				
Mean age	39.1	46.4	49.2	57.2
Mean number vascular disease	0.3	0.2	2.2	1.8
Mean number modifiable risk factors	3.9	2.4	4.2	2.6
Mean number non-modifiable risk factors	2.1	1.9	3.0	2.3
Number of health concerns	6.3	4.5	9.4	6.9
Lifestyle healthiness score	24.9	29.3	24.9	29.0
Proportion of population (%)	7.1	67.5	0.9	24.5
Two-step Solution 5				
Mean age (years)	43.4	44.6	46.7	57.5
Median age (years)	43	45	48	58
Mean number vascular disease	0.6	0.5	0.1	1.6
Mean number modifiable risk factors	3.4	3.5	2.0	2.3
Mean number non-modifiable risk factors	2.2	2.0	1.8	2.5
Number of health concerns	6.2	6.0	3.9	6.4
Lifestyle healthiness score	26.7	26.7	30.3	29.7
Proportion of population (%)	14.1	20.3	40.1	25.5

When the five variables identified by LCA were entered as categorical variables for two-step clustering, a five-group solution was produced with a silhouette co-efficient of 0.7, suggesting good cohesion and separation. When a four-group solution was forced, cohesion and separation remained good (silhouette co-efficient = 0.7). This solution had good internal consistency, as there was little change in the silhouette coefficient when the file was split by gender (0.6 for males and 0.7 for females) or by Air Miles status (0.6 for non-Air Miles and 0.7 for Air Miles participants).

The four-group solution (Two-step Solution 5) placed 40.1% in one group, with the remaining 59.9% of the population split into three groups of 25.5%, 20.3% and 14.1%. The overall large-to-small group size ratio was 2.84.

Proportions, means and group membership effect sizes for all variables are provided in Table 14 in Appendix 5. Of the variables not used for clustering, there was a large effect for age in years (ϖ=.378) and a medium-sized effect for medication use (Cramer's V_{1df} =.434). Of the clustering variables, the largest effect in Two-step Solution 5 was for high fat food consumption (Cramer's V_{1df}=.959), followed by high salt consumption (Cramer's V_{1df}=.908) and hypertension (Cramer's V_{1df}=.674). This was quite different from LCA Solution 2, in which the clustering variable with the largest effect size was hypertension (Cramer's V_{1df}=.984), followed by high fat foods (Cramer's V_{1df}=.664) and fast foods (Cramer's V_{1df}=.572). This suggests dietary behaviours had a larger effect on group membership in Two-Step Solution 5 than they did in LCA Solution 2.

In Two-step Solution 5, median ages of the four groups were 43, 45, 48 and 58 years. As shown in Table 18, mean number of vascular diseases increased with age

(which had a large effect of ϖ=.659). Mean number of modifiable risk factors had a large group effect (ϖ=.479) but did not, as anticipated, decrease with age. Even though effect sizes were large for all variables except lifestyle healthiness score (for which there was a medium-sized effect of ϖ=.131), differences between groups were small, making it difficult to determine how they could be distinguished from one another.

Figure 42: Proportions by Two-step Solution 5 groups for modifiable risk factors and vascular conditions

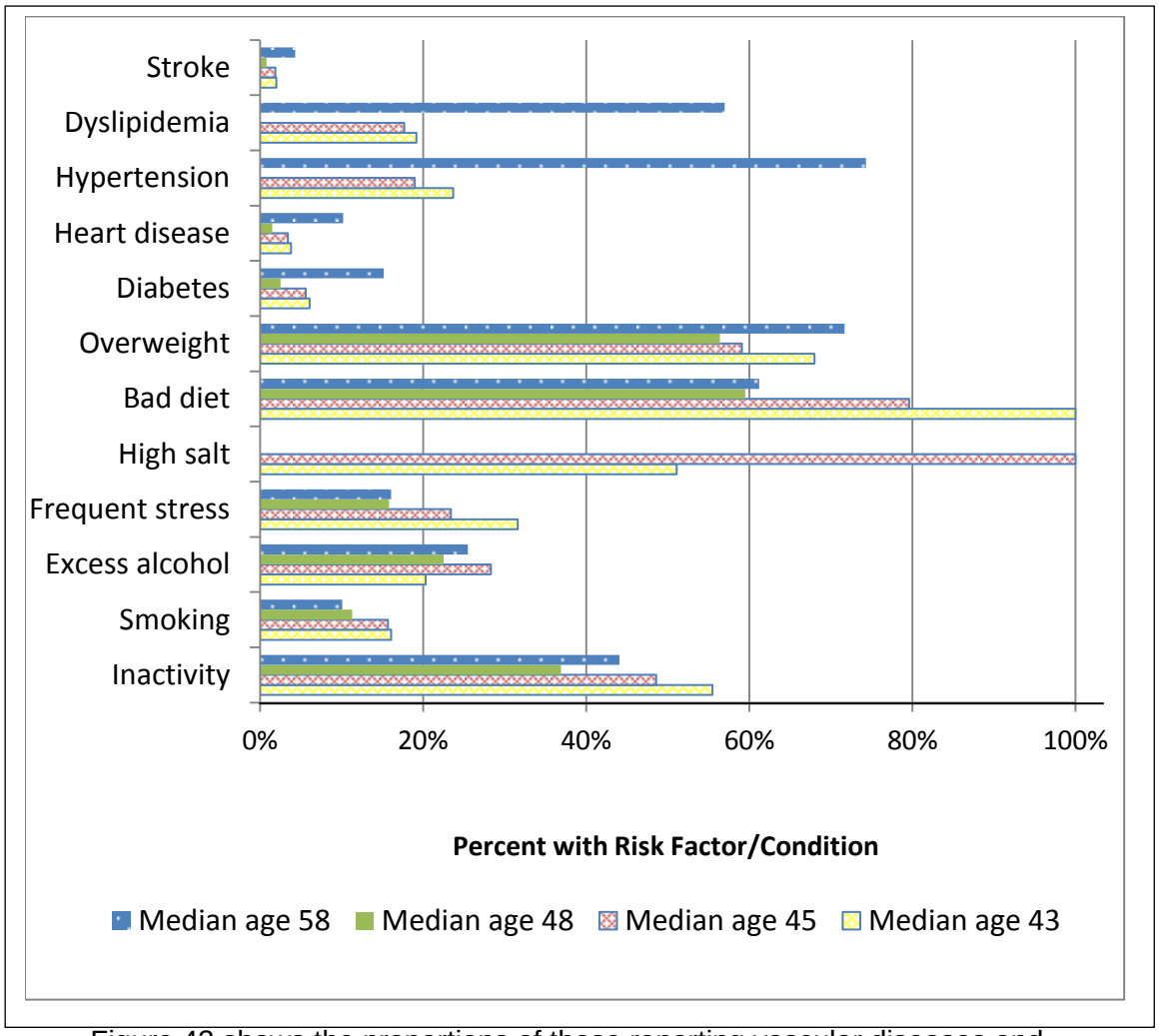


Figure 42 shows the proportions of those reporting vascular diseases and modifiable risk factors by Two-step Solution 5 group. The algorithm did not place any cases with stroke or dyslipidemia in the second-oldest age group (median age 48) and placed all cases of high salt consumption in the two youngest groups (100% of the group with the median age 45 and 51.1% of those in the median age 43 group). It was difficult to see any consistent patterns, particularly for modifiable risk factors. Effect of group membership on the distribution of non-clustering modifiable and non-modifiable risk factors, as well as the report of vascular diseases, readiness to change and health behaviours such as medication adherence, were consistently small (Cramer's V_{1df}<.300).

Table 19 gives an overview of the Two-step Solution 5 and LCA Solution 2. When numbered in order by age, agreement between solutions was poor (61,759/119,828 = 51.0%, Cohen's *kappa* = 0.266), suggesting groups formed by the different procedures were not similar. Of the two solutions, LCA Solution 2 appears to be less robust, as internal consistency and face validity were fair to poor. However, Two-Step Cluster 5 only had fair face validity, suggesting that it may not be optimal.

Table 19: Comparison of Segmentation Solutions: Fruit and Vegetable, Fish and Salt Consumption and Family history of Dyslipidemia or Premature Heart Disease as Clustering Variables

	K-means Clustering	**Two-Step Clustering (Two-step Solution 5)**	**Latent Class Analysis (LCA Solution 2)**
Potential solutions	None	Four- or five-group solutions are possible	
Internal consistency	N/A	Good	Poor
Group sizes of the selected solution	N/A	14% to 40%	Theoretically from 2% to 58%; When applied from 1% to 68%
Large-to-small group ratio	N/A	2.84	Theoretically 36.1; When applied 75.0
Face validity	N/A	Fair	Fair
Differentiation in 4-group solution	N/A	Large effect for age and medium-sized effect for medication use	Large effect for age, and medium-sized effect for family history of hypertension and medication use
Distance from clusters to center	N/A	N/A	N/A
Silhouette Coefficient (2-step only)	N/A	Good (0.7)	N/A
Classification error rate (LCA only)	N/A	N/A	13.6%
Agreement between 4-group solutions	N/A	51.0%, Cohen's *kappa* =0.266, *p*<.001	

Choosing an optimal solution

As noted by Romesburg (331), in exploratory cluster analysis, deciding the number of groups to be formed is not known and depends in large part upon the experience and purpose of the investigation. Although the decision to create four-group solutions was arbitrary, the fact that eight of the four-group solutions had good internal consistency suggests that this may be a reasonable approach.

Table 20: Comparison of Segmentation Solutions

Segmen-tation	K-means 1	Two step 1	K-means 2	Two step 2	K-means 3	Two step 3	LCA 1	Two-step 4	LCA 2	Two-step 5
Cluster-ing variables	# vascular diseases, # modifiable RFs, # non-modifiable RFs		# health concerns, lifestyle healthiness score		Age, lifestyle healthiness, # vascular diseases, # non-modifiable RFs		Binary fruit/veg, fish and salt intake, Hx dyslipidemia & heart disease		Binary high fat, fast food and salt consumption, Dx dyslipidemia, hypertension	
Internal consis-tency	+	+	+	+	+	+		+		+
Face validity			+	+						
Diff in non-clustering variable	Age +	Age ++	Age ++	Age ++	Med+ Marital + Employ +	Med +	Age ++	Age+	Age++ Med+ Family Hx HBP+	Age++ Med+
K-means cluster distance	++		++							
2-step silhouette co-efficient		+								
LCA error classsifca-tion error rate (%)							29.7		13.6	
Agree-ment (Cohen's *Kappa*)	0.367		0.602		0.288		0.565		0.266	

RF = risk factors Hx = Family history Dx = diagnosis Med=Poor medication adherence
Marital = married/common-law vs. not Employ = work full/part-time vs. not
+ Medium-sized effect ++ Large-sized effect

Table 20 summarizes the ten segmentations that were generated. It shows:

- two solutions had poor internal consistency: LCA Solutions 1 and 2
- eight solutions had poor face validity: K-means Solution 1 and 3, Two-step Solution 1, 3, 4 and 5, and LCA Solution 1 and 2
- the silhouette co-efficient suggested there was only fair cohesion and separation for four two-step solutions: Two-step Solutions 2, 3, 4 and 5
- the distance from cluster to centre had only a small effect size for one k-means solution: K-means Solution 3
- the error classification rate for LCA Solution 1 was large.

Based on this information, it can be said that of the ten solutions, K-means Solution 2 was the most robust and convincing. Furthermore, K-means Solution 2 provided a "narrative"

that helped to understand users by not only age but also readiness to make health behaviour change.

Summary

The analysis in this chapter showed the challenge posed by exploratory analysis of a data warehouse, i.e., data mining. Data are seldom uniform and even random fluctuations can be misinterpreted as patterns if statistical analysis is not performed in a thoughtful manner (373). Multiple segmentations must be conducted in order to generate at least one solution that is meaningful and has utility for program operators. To make the shift from "data dredging" or "data snooping" to true data mining or knowledge discovery in databases, solutions cannot be accepted blindly but must be considered and interpreted according to prior theory or knowledge about the sample population (249).

All segmentations created in this study suggested age was a primary and important variable differentiating HRA users. As described, K-means Solution 2 created groups that differed not only by age but, as lifestyle healthiness score was a clustering variable, by Prochaska's stage of readiness to change (251). It is possible that this quality of readiness to change may reflect internalized or intrinsic motivation for health-promoting behaviours (374) or is part of a greater health construct such as health information orientation (80), health consciousness (17, 20), and/or health conscientiousness (32). For example, Roberts *et al.* theorize that health conscientiousness increases during adulthood and midlife (32). Increasing health conscientiousness with age conforms to the negative relationship observed in the HRA data base between age and the prevalence of modifiable risk factors, as well as the positive relationships with variables such as readiness to change modifiable risk factors, medication adherence and control of medical conditions

Table 21 shows a breakdown of the HRA population by K-means Solution 2 Group. Two-thirds of participants could be described as "more healthy" in that they tended to have fewer modifiable risk factors and a greater readiness to change those they have. Messaging for such participants could, for example, focus on wellness and the prevention of relapse (251), perhaps by promoting feelings of autonomy (sense of control), competence and relatedness (374, 375) for health-related behaviours. Older participants made up the bulk of this group (61.4%), but a significant minority was younger. Presumably, the efficacy of messaging could be enhanced by not only tailoring to health needs but by targeting key demographics such as gender and life stage. For example, as noted in a National Institute for Health and Clinical Excellence (NICE) review, transition points in people's lives, such as entering the workforce, becoming a parent, or

retirement, are times at which people may be ready to review their behaviour and consider change (376).

Table 21: Breakdown of K-Means Solution 2 Groups by Age and Healthiness

<table>
<tr><th>By Group

n (% of HRA total population)</th><th>By Age

n (% of HRA total population)</th><th>By Healthiness

n (% of HRA total population)</th></tr>
<tr><td>Younger & Healthier
29,378 (24.4%)</td><td rowspan="2">Younger = 46,093 (38.2%)
Healthier – 63.7%
Less healthy- 36.3%</td><td>Healthier= 76,175 (63.2%)
Younger – 38.6%
Older – 61.4%</td></tr>
<tr><td>Younger & Less Healthy
16,715 (13.9%)</td><td rowspan="3">Less healthy = 42,766 (35.5%)
Younger =39.1%
Older = 60.8%</td></tr>
<tr><td>Older & Healthier
46,797 (38.8%)</td><td rowspan="2">Older = 72,848 (60.4%)
Healthier – 64.2%
Less healthy – 35.8%</td></tr>
<tr><td>Older & Less Healthy
26,051 (21.6%)</td></tr>
<tr><td>Missing (not clustered)
1,569 (1.3%)</td><td>Missing (not clustered)
1,569 (1.3%)</td><td>Missing (not clustered)
1,569 (1.3%)</td></tr>
<tr><td>Total
120,510 (100%)</td><td>Total
120,510 (100%)</td><td>Total
120,510 (100%)</td></tr>
</table>

A third of younger and older participants fell into the “less healthy” categories. Such individuals may respond better to messaging that takes into account not only life stage but the natural history of behaviour change and the fact that relapse is common (251), the age gradient of most chronic conditions, and the importance of building intrinsic motivation for change (375).

In conducting segmentations on databases generated by online users, health promotion and other agencies must proceed carefully and remember the acronym GIGO: “garbage in/garbage out.” Software can produce solutions fairly easily: the difficulty is creating and recognizing which solution or solutions is or are sound (i.e., reliable and valid) and useful for program purposes. Organizations may need to conduct several segmentations and compare them if they are to find the optimal solution for their population. It is possible that what was found in the HRA population may not occur in other online samples.

6: Predictive Ability of Groups Formed Through Segmentation

Segmentations are interesting and, in the case of freely available health etool populations, may help to guide message tailoring. But are they helpful in predicting online behaviour, such as enrolling for a follow-up etool?

To begin, it is helpful to understanding the follow up services. Users who completed an HRA had the option of enrolling for the follow-up email service (eSupport), the BPAP blood pressure management portal, or the HWAP. During the study period, 38.3% (n=46,197) of users enrolled for any of the follow up options (35.3% of males and 39.8% of females, Cramer's V_{1df}=.043).

The Air Miles incentive was designed in large part to halt the decline in enrollment for the email-based eSupport. When the email service was created in 2004, about half of HRA users enrolled but starting in 2008 there was a precipitous decline to as low as 2%. The reasons for this decline are unclear. There are three possible explanations. First, the decline may reflect a historical trend of growing consumer fatigue with emails and "email overload" (377, 378). Faced with ever-increasing volumes of email, consumers may become increasingly reluctant to enroll for an email-based service. Second, adding more options increased the decision-making burden on consumers (379), thus, in marketing terms, increasing the odds that consumers will "leave the store empty-handed" (380). The HRA system does not provide any recommendation agent or interactive decision aid to help consumers sort through their options or any comparison matrix to organize information or choices (381). Uncertain as to what option may be optimal, their commitment to change, and/or the burden that enrollment might entail, consumers may find it simpler to exit without making a selection.

Third, it is possible enrollment may be influenced by the proportion of HRA users who are, in the words of Cho *et al.* health information-oriented as opposed to behaviour-oriented (85). Finally, the three possible explanation are not necessarily mutually exclusive: one or more may be operational among HRA users.

Enrollment for follow up

Almost four out of ten participants (39.8% of women vs. 35.3% of men) enrolled for any form of follow up etool; the difference was statistically significant but the effect size was small (Cramer's V_{1df} =.043). There was no difference between men and women in the proportion who enrolled for eSupport (30.6% for both, Cramer's V_{1df} =.000, *p*=.967) but men were somewhat more likely to enroll for the blood pressure self-management

module (2.3% *vs.* 1.6%, Cramer's V_{1df} =.024) while women were more likely to enroll for the HWAP (9.2% *vs.* 3.2%, Cramer's V_{1df} =.108; both are small effects).

Table 22 shows enrollment by age group for the three etools. There appeared to be an inverse linear trend for eSupport but, as was the case for the BPAP and HWAP, the effect size was small.

Table 22: Enrollment for follow up by age group

Etool	18-34 yrs (%) n=23,052	35-44 yrs (%) n=21,758	45-54 yrs (%) n=31,226	55-64 yrs (%) n=29.098	65-74 yrs (%) n=12,463	75-90 yrs (%) n=2,931	Effect size (*eta*)
eSupport	32.6	32.9	29.9	29.9	28.3	20.7	.041
BPAP	0.7	1.5	2.1	2.3	2.3	2.4	.042
HWAP	6.7	7.3	8.4	7.6	6.0	3.7	.006
Joined any	39.4	40.9	38.9	37.8	34.8	25.8	.040

For *eta* (effect by interval), 0.01 = small effect, 0.06 = medium-sized effect, and 0.14= large effect; all comparisons significant at $p<.001$.

Education (Table 23) also had little or no influence on enrollment rates.

Table 23: Enrollment for follow up by education level

Etool	< High School	High School	Some Post-Secondary	University/College Graduate	Effect size (*eta*)
eSupport	27.0%	29.6%	31.3%	31.0%	.021
BPAP	2.3%	1.8%	1.7%	1.8%	.009 *
HWAP	6.2%	6.5%	7.1%	7.7%	.020
Any	34.1%	36.6%	38.8%	39.2%	.028

*p=.017; for all others $p<.001$
For Cramer's V1df, 0.01 = small effect, 0.30 = medium-sized effect, and 0.50 = large effect.

Entry portal had a large effect for the HWAP and eSupport, a medium effect for any enrollment, and a small-to-medium effect for the BPAP module, the least popular of the three etools (see Table 24). This suggests that to a large extent people come through the landing page that reflects their interest or need.

Table 24: Proportion who enroll by entry portal to HRA

Etool	H&S HRA	Mobile	BPAP	eSupport	HWAP	Total	Effect size (*eta*)
eSupport	4.1%	2.4%	0.0%	49.0%	3.3%	30.6%	.495
BPAP	3.9%	0.2%	0.1%	0.2%	0.1%	1.8%	.233
HWAP	6.6%	2.0%	0.0%	1.8%	44.7%	7.3%	.536
Any	13.8%	4.3%	9.1%	49.6%	44.7%	38.3%	.352

For Cramer's V1df, 0.01 = small effect, 0.30 = medium-sized effect, and 0.50 = large effect.

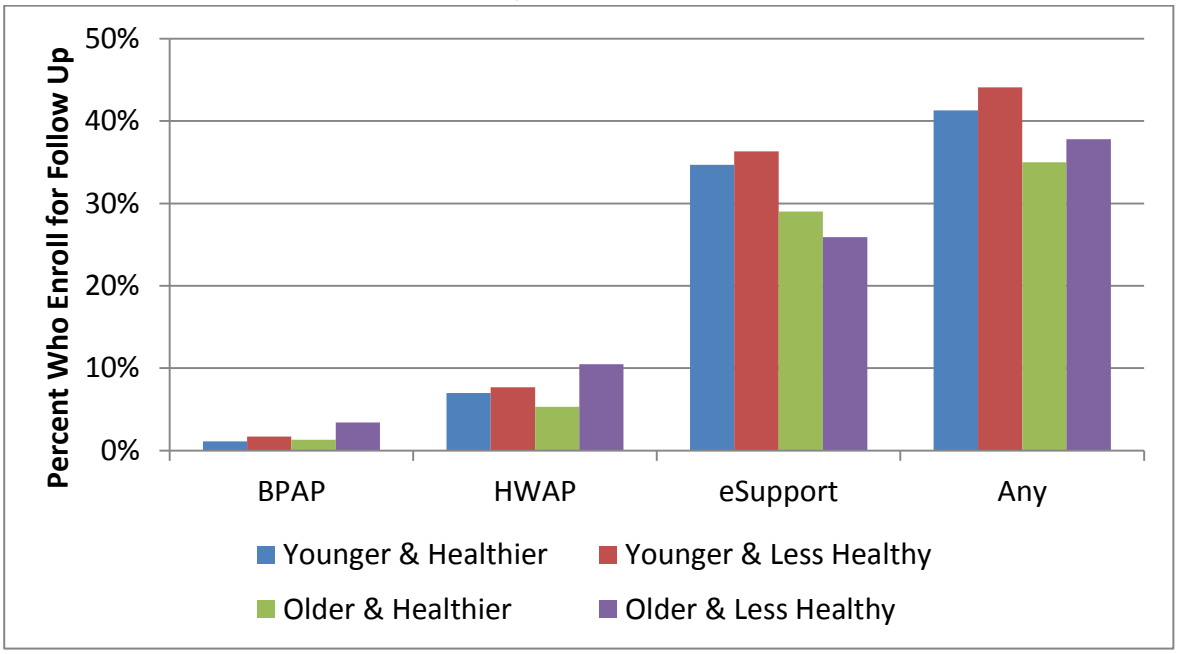
Although the number of non-Air Miles users who came to the HRA through the eSupport landing page was small (n=528), 40.7% enrolled for the service. This compares

favourably to the 49.0% enrollment rate of those who came to the eSupport portal through the Air Miles incentive. The difference between the two groups was statistically significant (p<.001) but the effect size was small (Cramer's V_{1df}=.014).

The follow-up etools address modifiable risk and since the k-means 2 segmentation is based on lifestyle healthiness and number of health concerns, presumably enrollment would vary by group. Testing whether there is a relationship would also help to determine whether the k-means group segmentation has practical or programmatic utility in that it predicts user behaviour.

Figure 43 shows enrollment by k-means group for the three follow-up etools as well as overall enrollment. For the BPAP and HWAP, the "less healthy" group of the two age dyads appeared to have higher enrollment rates than the "healthier" group. However, the effect sizes were small (respectively, Cramer's V_{1df}=.068 and .072). For eSupport, the pattern changed for the two older groups, with the "older and healthier" having a higher rate than the "older and less healthy" (29.0% *vs.* 25.9%). However, the effect size remained small (Cramer's V_{1df}=.082). Enrollment for any form of follow up had a similar trend to that observed for the HWAP and BPAP but again the effect size was small (Cramer's V_{1df}=.070).

Figure 43: Enrollment for follow up by k-means 2 group



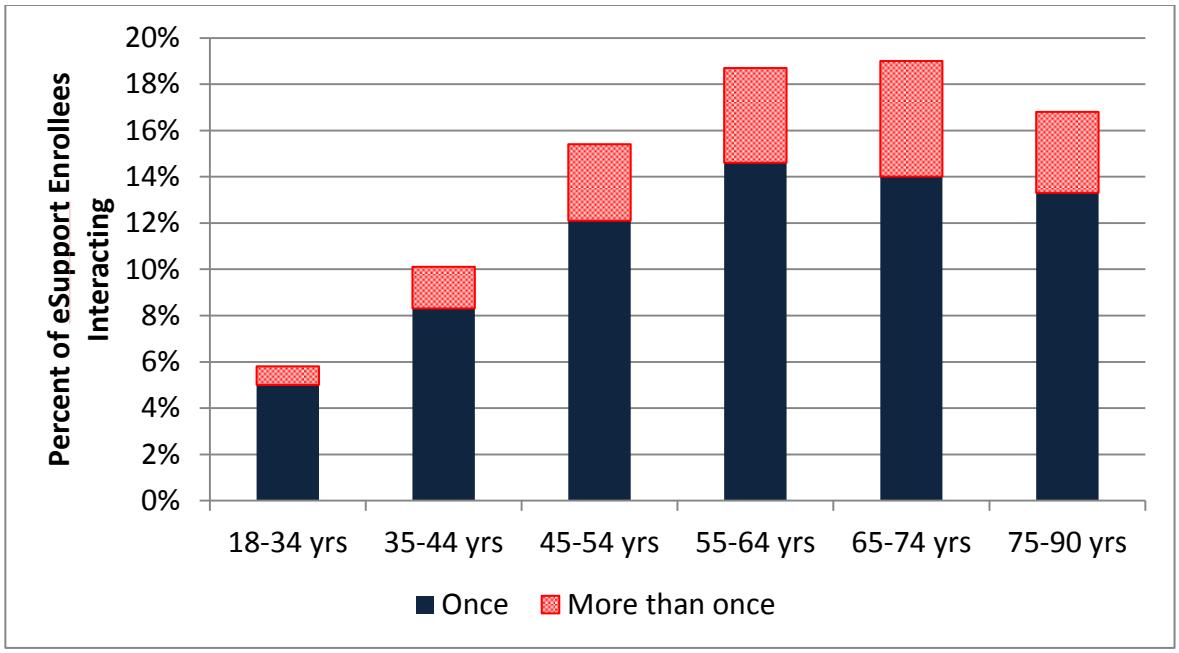
Interaction with the eSupport system

Of the 36,852 who enrolled for eSupport, the majority (31,698 or 86.4%) did not interact with the system, about 11% (3,942 or 10.7%) interacted once, and only 3% (1,060 or 2.9%) interacted more than once. There was little difference in the level of interaction by gender, with 87.5% of males and 85.9% of females failing to interact, 9.8% and 11.1% interacting once, and 2.7% and 3.0% interacting more than once (Cramer's V_{1df}=.022).

Once people enrolled, the Air Miles incentive had no effect on interacting with the system. The proportion of enrollees who interacted was non-significantly higher for non-Air Miles than Air Miles participants (16.2% *vs.* 13.5%, Cramer's V_{1df}=.015, *p*=.005). A similar pattern was observed when non-Air Miles and Air Miles participants were compared by whether they interacted once (12.3% *vs.* 10.6%) or more than once (3.8% *vs.* 2.8%). Effect size for this comparison was small (Cramer's V_{1df}=.016) and not statistically significant (*p*=.011).

Whereas enrollment for eSupport fell by age (see Table 22), there was a large effect (*eta*=.138) for interaction by enrollees to increase by age group (see Figure 44). The effect was not linear, however: for both interacting once and interacting more than once, rates were lowest for the 18-34 age group and then increased until the 55-64 age group and declined slightly for the 75-90 group. The category of interacting more than once was more skewed towards the older age groups than interacting only once (respectively, skewness =-.284, SE=.075 and -.193, SE=.039).

Figure 44: Interaction with eSupport among enrollees by age group



Logistic regression of ever interacting with eSupport by age group demonstrated the same trends observed in Figure 44 but for three of the age groups the effect was not statistically significant and the model explained only a small proportion of variance (see Table 25).

Table 25: Ever interacting with eSupport by age group

Age group*	β (se)	Wald	*df*	*p*	OR	95% CI
34-44 yrs	-1.183 (.120)	97.732	1	.000	.306	.242-.387
45-54 yrs	-.590 (.116)	23.905	1	.000	.554	.442-.696
55-64 yrs	-.094 (.113)	.708	1	.400	.910	.729-1.134
65-74 yrs	.149 (.117)	1.447	1	.229	1.145	.918-1.427
75-90 yrs	-1.603 (.109)	1.622	1	.203	1.161	.923-1.461

* Compares enrollee who never interacted to those who interacted; reference category = 18-34 years
β(se)=beta co-efficient (standard error) OR=Exp(β) followed by lower and upper 95% CI

R^2: 0.023 (Cox and Snell), 0.042 (Nagelkerke); Model X^2<.001; Hosmer and Lemeshow X^2 =1.000

Table 26 shows interaction by k-means 2 group membership. The two younger groups had similar rates of ever interacting (15.3% for the "younger and healthy" and 15.6% for the "younger and less healthy") but the two older groups varied, with the "older and healthier" having the lowest proportion interacting with the system (7.7%) and the "older and less healthy" the highest proportion (21.0%). Note that although the "older and less healthy" group had the lowest enrollment rate (25.9%), it paradoxically had the highest rate of ever interacting (21.0%). However, k-means group had only a small effect on the proportion of enrollees who interacted (Cramer's V_{1df}=.145).

Table 26: Interaction of eSupport enrollees by K-means 2 group membership

K-means 2 Group	eSupport enrollment n (%)	% of enrollees who interact		
		Once n (%)	>Once n (%)	Total n (%)
Younger & healthier	19,181 (34.7)	1,261 (12.4)	300 (2.9)	1,561 (15.3)
Younger & less healthy	6,067 (36.3)	718 (11.8)	231 (3.8)	949 (15.6)
Older & healthier	13,590 (29.0)	836 (6.2)	205 (1.5)	1,041 (7.7)
Older & less healthy	6,735 (25.9)	1,097 (16.3)	318 (4.7)	1,415 (21.0)
Total	36,573 (100)	3,912 (10.7)	1,054 (2.9)	4,966 (13.6)

Although k-means group membership had only a small effect, multinomial regression was conducted to see if it could help explain eSupport interaction by enrollees (see Table 27). Compared to the reference category of the "younger and healthier," there was no statistically significant difference in the odds of the "younger and less healthy" interacting once (OR=.96, 95% CI .87-1.06, *p*=.402) but there was a modest and significant effect for interacting more than once (OR=1.30, 95% CI 1.09-1.56). Compared

to the “younger and healthier,” the “older and healthier” were less likely to interact once (OR=0.46, 95% CI 0.42-0.50) or more than once (OR=0.47, 95% CI 0.40-0.57). In contrast, the “older and less healthy” enrollees were 40% more likely to interact once (OR=1.41, 95% CI 1.29-1.54) and 73% more likely to interact more than once (OR=1.73, 95% CI 1.47-2.03).

Table 27: Comparison of eSupport enrollees who never interacted to those who interacted once or more than once

K-mean 2 group*	β (se)	Wald	*df*	*p*	OR	95% CI
Interacted once:						
Younger & less healthy	-.042 (.050)	.702	1	.402	.959	.870-1.058
Older & healthier	-.787 (.047)	283.181	1	.000	.455	.416-.499
Older & less healthy	.343 (.045)	258.671	1	.000	1.410	1.291-1.539
Interacted > once:						
Younger & less healthy	.266 (.090)	8.794	1	.003	1.304	1.094-1.555
Older & healthier	-.746 (.092)	65.981	1	.000	.474	.396-.568
Older & less healthy	.548 (.083)	43.962	1	.000	1.729	1.471-2.033

* Compares enrollees who never interacted to those who interacted once or more than once; reference group is “Younger and healthier”
β(se)=beta co-efficient (standard error) OR=Exp(β) followed by lower and upper 95% CI

R^2: 0.022 (Cox and Snell), 0.036 (Nagelkerke), 0.024 (McFadden); Model X^2<.001.

Although results shown in Table 27 confirm the trends observed, k-means 2 group membership explained only a small amount of variance. Adding age as a co-variant increased the amount of variance explained to a modest 4% but would be confounded by the relationship between it and k-means 2 groups (data not shown).

Change in readiness

Of the 36,852 who enrolled for eSupport, 5,154 (14.0%) chose a focus area. Stage of change was either carried over from the self-report in the HRA or edited by the user. On a scale ranging from one (Precontemplation) to four (Action), mean readiness to change initially averaged 2.92 (2.9).

Of these 5,154 users, 1,102 (21.4%) returned and rescored their readiness of change. Readiness to change at this point averaged 2.97 (3.0), a difference that was not statistically significant (*p*=.093) when tested using repeated measures ANOVA.

The proportions reporting change varied according to their initial stage of change. Thus:

- of 221 who started in the Precontemplation stage, 154 (69.7%) stayed in that stage and 67 (30.3%) progressed to a greater readiness to change;

- of 95 who started in the Contemplation stage, 5 (5.3%) regressed, 54 (56.8%) remained the same, and 36 (37.9%) progressed;
- of 334 who started in the Preparation stage, 30 (9.0%) regressed, 224 (67.1%) remained the same, and 80 (24.0%) progressed; and
- of 452 who started in the Action stage, 345 (76.3%) remained the same and 107 (23.7%) regressed.

These findings suggest that eSupport emails may have a small, positive effect on preventing relapse and increasing the readiness to change, particularly for those beginning in the Precontemplation, Contemplation or Preparation stages. For those in the Action stage of change, it is possible that eSupport may be helpful in preventing relapse. However, due to the small number of people who rescored themselves and the lack of statistical significance, these results must be interpreted with caution.

Summary

One of the objectives of this chapter was to see if the groups formed through segmentation had utility in predicting the behaviour of HRA users. Although some weak trends were observed, the segmentation had virtually no predictive ability. As well, neither age group nor SES as represented by highest level of education explain enrollment.

The majority of people who enrolled for eSupport accessed the HRA through the eSupport landing page. The Air Miles incentive increased the number of people coming through the eSupport portal but had no significant effect on whether those who came through this landing page ended up enrolling for the email service.

Only a minority (14%) of those who enrolled for eSupport interacted with the system. Interaction was not influenced by the Air Miles incentive but -- unlike enrollment – there was a large effect for age group. The relationship was not linear and in logistic regression had poor explanatory power. Similarly, although k-means 2 group membership appeared to influence interaction with the system, the effect was also small.

The failure of the k-means 2 segmentation to explain variance in either enrollment or etool interaction means the grouping may be useful in tailoring health assessment information but has no utility in predicting follow-up etool behaviour. Information other than gender, age, education or k-means 2 group may be required to understand what separates the majority of enrollees who do not engage with the email service from the minority who do.

Understanding more about what triggers etool interaction could be helpful. Currently, only a small proportion of users interact but of those who do there is limited evidence the system may be helpful in promoting readiness to change or preventing relapse. Given the small number of people who interact, these results must be interpreted with caution. However, experimental research on the eSupport system have provided some evidence of efficacy for a clinical sub-set of the population, those with hypertension (254). More research could be helpful in exploring what conditions promote not just health information seeking but behaviour change and eSupport interaction.

.

9: Discussion and Conclusions

This chapter will attempt to pull together the results of the various analyses and discuss their implication for electronic health promotion, as well as organizations operating health etools.

Health information seeking is a common activity

In Canada, the U.S. and the U.K., approximately 80% of the adult population is using the Internet (3, 5, 6). One of the most common online activities of adults is searching for health information (51, 53, 55), with surveys estimating the proportion to range from 70% in Europe to 80% in the U.S. (348). The Pew Center estimates that of the entire American adult population, about 60% are Internet health information seekers (51). Health information may be of particular importance to the aging "Baby Boom" generation, which represents 30% of the population of Canada (382) and 44% in the U.S. (383).

Between 2004 and 2011, approximately 777,000 HRAs were completed and 178,000 users registered for the email-based eSupport system, with 98% of users based in Canada. Since the Canadian population of Internet health information seekers may total 16.2 million (60% of the 27.1 million adults aged 20 and over), it can be inferred that the HSF etools have yet to saturate the market and there is still potential for significant growth. Growth may depend in large part upon the organization's ability to market and advertise the site in order to turn the considerable pool of Internet health information seekers into more activated health etool users. As shown in Figure 1 in Chapter 3, traffic volumes fluctuates greatly over time and is largely dependent upon promotional activities.

HRA users are a distinct sub-set of the general population

HRA users self-select early in the online process. Of all visitors to the HRA landing page, approximately half left without starting the questionnaire. Those who started appeared to have a high level of motivation: up to 80% completed the HRA, even though it can take 12 or more minutes to complete.

Previous research has shown that Internet health information seekers are not representative of the general population. Rather, they tend to be younger, as opposed to older, adults (48, 49, 58, 59, 61, 62, 65, 69), female rather than male (48, 49, 61, 65, 73,

74), more highly educated (38, 48, 55, 56, 58, 62, 67-69, 72-74), motivated by either medical issues or poor health (48, 49, 58, 61, 63, 64, 68, 70, 73, 74, 78), or be those with good health (62, 66, 77) seeking information for themselves because they are health-oriented (76-78) or on behalf of family members or friends (61, 68, 70). Some of this skewing is undoubtedly due to the "digital divide" still affecting Internet use (54, 57, 384).

The HRA population conformed to these descriptions of health information seekers, being skewed towards females (69.0% of participants), adults between the ages of 45 and 64 years (50.0%), university or college graduates (60.0%), those who are married (58.3%), and with one or more modifiable lifestyle-associated risk factor (94.4%) or a vascular disease (38.7%). Direct comparison of Canadian HRA participants to the general population of Canada using odds ratios confirmed that the HRA significantly over-represented some segments of the population but under-represented others. Over-represented segments included women and those aged 45 to 64 years, more highly educated individuals, or those reporting hypertension, overweight/obesity, asthma or mood disorders. Under-represented segments included men, smokers, the youngest (aged 20 to 34) or oldest (65 to 89 years) age groups, those with COPD, and older adults with diabetes or arthritis. Furthermore, it was shown that in the case of education, weighting the HRA sample by age and gender could not remove the large and statistically significant differences between what was observed (the HRA sample) and what would be expected from national statistics.

In some respects, such as the skewing by gender, age and education, the HRA population resembles samples recruited for health etool RCTs. Does this mean RCT samples may be representative or appropriate proxies for users of open-access, freely-available etools? Comparison with a convenience sample of three RCTs suggest that for the HRA, the answer is no. Even when the HRA population was restricted to more closely resemble RCT-recruited samples (e.g., by age or work status), there were significant differences. Analysis of the HRA using the natural experiment provided by the Air Miles promotion suggests that incentives, which may be used by some RCTs to improve recruitment or retention, may increase the number of users but not necessarily change the demographics or health profile of users of open-access etools, although the effect in RCTs has yet to be fully described. Finally, at this point, due to the paucity of published research on open-access etools, it is unclear whether user populations are similar or differ from one another. Limited data from a recent publication concerning the Heart Age calculator (232) suggests users may differ from those attracted to the HRA; however, these findings are tentative and no firm conclusions can be drawn. If there is variance in open-access etool populations, then the task of comparing them to RCT samples would become more complex. As a result, the best evidence for understanding the users of

open-access etools probably comes from the etools themselves, rather than other, proxy samples.

An important implication of finding systematic bias in the HRA sample is that it gives insight into which segments of the population are unlikely to utilize the etool. These may include adults with only limited education who, because of the social determinants of health, are at increased risk of chronic disease (385, 386), and those who place a low value on maintaining or improving their health (96) and/or are health information avoidant (35). Psychodemographic analysis by Navarro and Wilkins in 2001(233), for example, found "almost 40% of today's health care consumers are not inclined to seek health information on the Web, or anywhere else for that matter" (pg. 8). Such people may place a low value on health information, have a low propensity to be proactive about health care, or distrust medical information or professionals (233). As such, they are unlikely to participate in health etool RCTs (204) or to use open-access health etools (98).

In studying Internet health information seeking behaviour, many studies have utilized theoretical approaches in which rational and active choice play important roles, such as the Theory of Planed Behaviour, Technology Acceptance Model, Health Belief Model, or Uses and Gratifications Theory (387). As noted by Marton and Choo such models are not particularly good at incorporating elements such as emotions or social determinants (387). Without understanding the psychological or social subtext of users' lives, such theories may not be helpful in understanding people such as Navarro and Wilkins' "Clinic Cynics" or "Avoiders" (233), Miller's information "blunters" (37), or those of lower socioeconomic status (388). For example, even among those diagnosed with chronic disease, some may see themselves as having so little control or agency they do not see health information as beneficial (389, 390). Such people may lack the capacity, motivation or skill to use their personal power to become what Archer (2007) refers to as "active agents" (391). Organizations operating health etools need to consider the implications of these factors for the effectiveness and cost-efficiency of health promotion efforts (230).

HRA users are not homogeneous

As discussed by Evans, "the broadest approach to audience segmentation is targeted communications, in which information about population groups is used to prepare messages that draw attention to a generic message" (392). Kreuter *et al*. (102) note that targeting is based upon the principle, borrowed from marketing, of market segmentation, i.e., that "sufficient homogeneity exists among members of a demographically defined population to justify using one common approach to communicate with all of its members"

(pg. 5). However, variance occurs even within demographic groups. For example, Gallant and Dorn's study found the most consistent and powerful predictor of preventative health behaviours was not age, gender or race/ethnicity but baseline behaviour (393).

Because of the limitations of broad demographic categories, in social marketing increasing attention has focused upon a "a more specific, individualized form of segmentation" referred to as tailoring (392). Tailoring is distinguished by the fact that messages are based on individual-level factors such as behaviours, needs, or attitudes (394). Such messages are then directed towards specific individuals within organizations or demographic groups (322, 394). By more accurately targeting the needs, attitudes or interests that cross demographic categories, tailored messaging are more individualized and as a result may be more effective in capturing and holding the attention of users (328). The computational capacity of the Internet has greatly expanded the ability of health promotion/education materials to be tailored to the specific needs of users, while still preserving privacy or at least the illusion of privacy (12).

The effectiveness of tailoring depends in large part upon the validity of program developers' or operators' understanding of the characteristics of the intended audience. Without evidence from analysis of actual users, decision makers are often forced to fall back upon assumptions about the characteristics of future or current users and hence their perceived needs. For example, in a report to the HSFO, external, academic-based ehealth experts advised that HRA messaging should be tailored to match users' type of disease (heart disease or stroke) or primary concern (prevention or disease management) (personal communication with HSF managers). This study's analysis of the HRA population suggests that such approaches would be of only limited utility, as only 4.4% of the HRA population reported a diagnosis of heart disease and 2.1% stroke; moreover, although 61.3% of users may be described as "prevention oriented" because they do not report any current vascular disease diagnoses, almost 40% (37.1%) have both vascular diseases and modifiable risk factors.

As more social marketing segmentations of health consumers become available in the public domain, it becomes increasingly obvious that these populations are not homogenous or monolithic. For example, a segmentation of digital health consumers developed in 2014 by the American market research firm Park Associates reported four segments varying by health status and level of health consciousness (395). Two of the four were free of chronic health problems but varied in their level of health consciousness, one being called the Healthy and Engaged (26% of total) and the other Young and Indifferent (21%) (395). The other two categories included those who had been diagnosed with at least one form of chronic condition, of which one was health conscious

(Challenged but Mindful, 25% of total) and the other not (Unhealthy and In Denial, 28% of total) (395). It should be noted that in some respects this segmentation resembles the k-means 2 solution developed in this study.

It can be hypothesized that HRA users are individuals in whom health concerns have engendered not only health consciousness (20), a concern about health and well-being, but a health information orientation (35) which is evidence by health information seeking behaviour (13). At the very least, because they have completed the HRA, all users in this sample can be described as Internet health information seekers. It is possible, but beyond the scope of this study, that this is only one form of health-information seeking behaviour undertaken by users (80). Moreover, even though all HRA users may be assumed to be health-information seekers or to be health oriented, it is unlikely a population this large would be homogeneous. Indeed, the analysis conducted for this study suggests that although HRA users are health oriented they, like the Parks Associates' segments (395), vary in the degree to which their behaviour reflect health conscientiousness.

Grouping health etool users by need or attitudes, segmentation has the potential to provide more useful sub-sets for message tailoring (317, 318, 395). As noted above, tailored messages are not only more effective in capturing and holding the attention of users but may be perceived as more convincing (322, 328). As a result, this approach is being used by a variety of for-profit and not-for-profit organizations. For example, the U.S. market research firm Deloitte used factor and cluster analysis to develop six health consumer segments: 1) Casual and Cautious, 2) Content and Compliant, 3) Online and Onboard, 4) Sick and Savvy, 5) Out and About and 6) Shop to Save (326). Age and gender varied somewhat between groups but demographic categories such as age groups had unique profiles in terms of their proportions belonging to different segments (326). The value of the segmentation lay in its ability to identify individuals within demographic strata who were more, or less, likely to be pro-active in managing their health insurance needs or adhering to treatment (326), information of great interest to health insurers.

Segmentation can also be used to predict groups of potential users. In November, 2014, a webinar sponsored by Public Health Ontario shared information about the development of a new online cancer risk assessment website by the provincial government agency, Cancer Care Ontario (396). As described during the webinar, polling conducted during the developmental phase identified those segments of the Ontario adult population which were the most likely to utilize this type of resource (i.e., "early adapters") as well as those who might be at greater health risk but be less likely to visit (396). Four

of the five segments were predominantly 35 to 54 years of age (two early adapters and two of the less likely to visit groups) and three of the five were largely female (two of the early adapters and one of the less likely to visit). Rather than demographics, the key characteristics distinguishing segments and their likely response to the tool were healthy literacy and attitudes about disease prevention (396). Reflecting the analysis of Rimer and Kreuter (328), this information was used by the organization to guide the development of the look, tone, content and readability of the site so as to optimize the likelihood of capturing and retaining the attention of its most likely user segments. Moreover, as will be discussed later, understanding segments helped the agency to develop strategies to optimize reach among the at-risk but less-likely-to-visit segments (396)

Segmentation of a user data base is a complex process

Exploratory segmentation can be a helpful tool in giving new, and sometimes unexpected, insights into a population, particularly when dealing with the large volume of cases that can be generated by etools. However, segmentation must be informed by a solid understanding of the population of users, how data were collected (e.g., how questions were phrased), the different types of segmentation procedures, and other activities (e.g., promotions or incentives) that may have influenced the number or type of participants. As noted by Jain *et al*., in data mining it is relatively simple to extract information: the challenge is to extract meaningful information (397).

Four challenges were encountered in segmenting the HRA data base. First was choosing which segmentation procedure to use. It is natural for those analyzing data bases such as the HRA to wonder which segmentation procedure is “better” to use and although there are publications comparing the validity of different approaches they tend to be highly technical and often discuss techniques not easily accessed by businesses through common statistical packages such as SPSS (398, 399). As a result, such discussions may not be appropriate for organizations such as not-for-profits that are trying to manage online programs in a cost-efficient manner using commercial, standardized software.

As a general rule, due to the complexity of the calculations involved in hierarchical clustering, this procedure is not appropriate for the analysis of large data bases (335). K-means clustering is available in a number of commercial statistical packages, is capable of handling large databases, and is widely used in research. Perhaps its greatest limitation in exploratory research lies in the need to specify the number of clusters

(something that may be unknown), followed by its inability to handle categorical variables (335, 341, 397). The two-step clustering procedure included in the base SPSS package addresses some of the problems posed by the k-means procedure, in that it can combine categorical and continuous data and will suggest the appropriate number of clusters (344). In addition, the two-step procedures provides an estimate, the silhouette co-efficient, of the degree of cohesion within groups and separation between them, which can suggest the robustness of a solution.

LCA may be appropriate if there is reason to suspect that variables used for clustering are independent of one another but influenced by a hidden or latent factor (400). Moreover, there is a reasonably-priced commercial package (LatentGOLD) that can handle various types of data and gives considerable information on the goodness of a model fit (including the BIC, L^2, BVRs and error classification rate). In the case of the HRA, LCA was not particularly helpful: of five combinations of clustering variables, statistically significant solutions could be generated by only two. Furthermore, neither of the two LCA solutions was particularly robust or added more to the understanding of the HRA population than what was achieved using k-means and/or two-step clustering.

The second challenge was choosing which variables to use as clustering factors. Within a data base as large as the HRA, there are an almost infinite number of variables and combinations of variables that could be used for clustering. In data mining, particularly big data analysis, the impression is sometimes given that with enough data patterns will emerge even though the activities themselves appear to be unconnected. This is, in fact, the underlying premise of LCA: observable variables may have no connection to one another other than the fact that they are connected to an unobserved (latent) construct (401). However in large data bases even random fluctuations can give the appearance of patterns (373); as a result, the quality of a segmentation is dependent upon the care chosen in selecting clustering variables (315). Understanding the variables and the care taken in selecting those to be used for clustering is what separates “data dredging” from “knowledge discovery in databases” (249).

In the case of the HRA, although many combinations could generate statistically significant solutions, groups were more meaningful when the clustering variables were more highly correlated with one another and included information on readiness to change. When variables were not highly correlated, groups tended to form based largely on age, rather than health needs.

The third challenge in this exploratory segmentation was deciding on the number of segments. As described earlier, in exploratory segmentation there may be little or no evidence upon which to determine the optimal number of clusters (331). In the analysis of

the HRA, the number of clusters was arbitrarily set at four, although three- or five-group solutions were often statistically significant. A smaller number of groups may be convenient for marketing purposes but it must be acknowledged that there is no "correct" number. Navarro and Wilkins, for example, divided American health care consumers into nine groups (233) while an examination of Canadian adults identified 12 "value tribes" distinguished by attitudes, lifestyles and age (239).

The number of groups in exploratory segmentation is largely arbitrary as these are not "true" groups (intransitive reality) but representations created through the retrospective analysis of observable data (empirical events) (247). As Dolnicar (2003; pg. 10-11) puts it:

> Rarely is there an empirical data set where individuals form homogeneous groups that are clearly and distinctly separated from other homogeneous groups. Consumer heterogeneity is an individual phenomenon. As such, all grey shades exist and most groupings of such individuals into market segments represent an artificial task. Market segments are constructed, not revealed (333).

Finally, the fourth challenge was choosing which of the ten segmentations or solutions would be optimal for the HRA population. In this study, both objective (based on quantitative measures) and subjective methods of evaluating solutions were used. The primary objective approach was looking for substantive differences between groups in variables not used for clustering (332, 340) using effect sizes (297). Other objective measures included looking at internal consistency or reproducibility of the segmentation when the file was split, distance from clusters to centre (k-means), silhouette co-efficients (2-step) and classification error rate (LCA).

The primary subjective method for evaluating solutions was to consider the face validity of solutions, i.e., the extent to which they appeared to reflect patterns previously observed in the population (in this case, age-related trends) (231). This approach was only possible because of prior descriptive analysis of the data base (249).

As discussed by Harle *et al.*, program operators need segmentations that address operational needs by giving them useful insights into their client populations (229). A segmentation that distinguishes solely by age, for example, provides nothing more than what could be obtained through descriptive statistics. Other considerations important for program operators may include the number and relative sizes of the groups (e.g., writing tailored messages for a large number of small groups may not be cost-efficient), as well as the consistency of the groups (e.g., do they reappear when the file is split or new cases are added?).

Choosing the optimal solution for the HRA population

As described in Chapter 7, using different procedures and combinations of clustering variables produced ten segmentations. Of the ten, one, K-means Solution 2, appeared to be the most robust, in that it had good internal consistency and face validity, a large effect size for the distance from cluster to centre, and at least a medium effect size for the non-clustering variable of age. Furthermore, as the clustering variable of lifestyle healthiness score incorporates readiness to change, groups formed in this solution varied by their vascular disease burden, CVD risk factor burden and readiness to modify lifestyle risk factors.

In this segmentation, two groups of the same general age tended to be differentiated from one another in a way that suggested they were either "healthier" or "less healthy" for people of that age strata. There were trends for the "healthier" groups of each dyad to report better adherence with prescription medication regimes and, if applicable, blood pressure or blood glucose control. In other words, those in the "healthier" group may share a generalized set of personality traits characterized by health consciousness, health orientation, and a tendency to engage in behaviours that are health-promoting (32). In contrast, compared to their similar-aged peer group, those in the "less healthy" groups had more vascular disease and modifiable risk factors and poorer rates of control over medical conditions such as diabetes or hypertension. Such individuals may be interested in acquiring health information, as evidenced by their completion of an online HRA, but lack a health maintenance motivation sufficient to support health-enhancing behaviours (82). As noted above, in terms of variation in health consciousness, this grouping is somewhat similar to the Parks Associates' segmentation of digital health consumers (395).

To summarize, the four groups formed by K-means Solution 2 consisted of:

1) Younger and Healthier: This group had the lowest mean number of vascular diseases, which conforms to the general trend for the prevalence of vascular diseases to have a negative relationship with age. Compared to its less-healthy peer group (Younger and Less Healthy), this group had a lower mean number of modifiable and non-modifiable risk factors, were less likely to have poor medication adherence, and, for those diagnosed with hypertension or diabetes, to more frequently keep their condition in a healthy range. Overall, the picture emerged of a group that was relatively healthy and more health conscientious than its similar-aged peer group, although less health conscientious than some older participants.

2) Younger and Less Healthy: Of the four groups, this group had the lowest proportion of participants age ≥ 65 years (6.4% vs. 7.4% for the Younger and Healthier group, increasing to ≥15% for the two older groups). Of the four groups, it had the highest mean number of modifiable risk factors, the highest proportion with poor medication adherence, and the lowest proportion of good blood glucose and blood pressure control. It also had the second-highest mean number of vascular diseases, second only to the Older and Less Healthy group. As a result, this group could be nick-named the "The Young and the Careless," for they appeared to be the least health conscientious of the four.
3) Older and Healthier: The group had a median age five to six years older than the two younger groups (51 years) and 15% of participants were age ≥65 years. Compared to the other older group, this group had a lower mean number of vascular diseases and of modifiable and non-modifiable risk factors. Perhaps the most striking aspect of this group was that, of all four groups, it had the lowest mean number of modifiable risk factors, the lowest proportion who missed their medication some or most of the time and, for those diagnosed with hypertension or diabetes, the highest levels of good control. In other words, this group consisted of highly health conscientious older adults.
4) Older and Less Healthy: This group had the oldest median age (56) and the highest proportion (19%) of users aged ≥65 years. Of the four groups, it had the highest mean number of vascular diseases and non-modifiable risk factors. Its mean number of modifiable risk factors was greater than the other older group but less than the two younger groups, which fits the trend in the HRA for the number of modifiable risk factors to have an inverse relationship with age. The level of poor medication adherence did not vary from that of the Older and Healthier and the Younger and Healthier groups (respectively, 11% *vs.* 10% and 11%) but less than the Younger and Less Healthy (30%). Similarly, the proportions of those with hypertension or diabetes who kept their readings in a healthy range were less than the Older and Healthier group but greater than either of the two younger groups.

This segmentation fits the age-related trends already observed in the data and formed a narrative that made intuitive "sense" in understanding the large and diverse HRA population. It recognizes that health information seekers vary by the type of information they want, their motivation, and what they intend to do with the information (78, 85, 402-404). Not all consumers who seek health information are necessarily motivated or ready to modify their behaviour (89).

It should be noted that the analysis conducted in this study was a *post hoc* segmentation based on analysis of existing cases, rather than an *a priori* segmentation in which cases are classified as they enter the system according to pre-selected variables (333). The advantage of *a priori* segmentation is that it can be applied prospectively to users as they enter the system and thus be used to select tailored messaging. However, the limitation of *a priori* segmentation lies in the fact that classification is not based on evidence. In contrast, a *post hoc* segmentation is based upon analysis of empirical data; because it is retrospective, however, it may not be sensitive to changes over time in the population.

If groups from *post hoc* segmentations are merely retrospective statistical constructs, what practical use are they to organizations operating online health etools? There are two answers. First, such groups provide a way of helping organizations to think about their user population and see sub-groups that may not be readily apparent. This information can then be used to further refine messaging, guide marketing activities or shape health promotion strategies. Second, learnings from *post hoc* analysis may be used to develop selection or classification variables for prospective, *a priori* segmentations. For example, polling firms frequently use complex sets of IF statement logic to re-create psychodemographic segments identified through *post hoc* analysis of an earlier poll. Another option may be to use the *post hoc* categories as "archetypes" and then to weight cases on the degree to which they conform to them (344, 405). In the case of the HRA, for example, this could involve developing "Younger and Healthier," "Younger and Less Healthy," "Older and Healthier" and "Older and Less Healthy" archetypes and then constructing a weighting system so cases can be categorized as HRAs are completed according to which archetype they most strongly resembles

Why are some users not ready for change?

The very act of coming to and completing the HRA demonstrated a health information orientation on the part of users. The k-means 2 segmentation suggests that even though all users were health information seekers, about a third (36%) had low readiness to follow or adopt the sort of behaviours that are health-enhancing. This finding is not surprising, given the analysis of Moorman and Matullich which suggests information acquisition and health maintenance represent different and not necessarily overlapping motivators for HSIB (82). This raises the question: how do people – even those who are health-information-oriented – make the transition between good intentions to health-enhancing behaviour?

Although the Transtheoretical Model of Change was utilized in the HRA, as it is for many health etools (105), it may not be particularly powerful in explaining the process by which people are activated. A more fruitful approach may be provided by the Health Action Process Approach. In this theory, change is viewed as consisting of two processes: intentional and volitional (406). The development of intention (the intentional stage or process) is influenced by three factors: the user's outcomes expectancies for change, perceived self-efficacy, and risk awareness (407). While risk awareness is the predominant theme of the HRA, it has been described of the weakest of the three influences on intention development (407), perhaps because it stems from external forces (408). For the HRA, driving the development of intention may require a greater focus upon building users' outcome expectancies and perceived self-efficacy.

The second stage or process in the Health Action Process Approach is named "volitional" and consists of two types of people: those who intend to change but haven't yet started and those who have been able to start translating intentions into action (407). Further research might be helpful to see if this distinction can be applied to the HRA population. Are the "Younger and Less Healthy," for example, those who are seeking health information because they would like to be healthier but lack sufficient motivation to actualize those intentions? Unfortunately, because the HRA data base is observational and retrospective, it lacks the variables required to study these sorts of processes or relationships. Further research would require the co-operation of the HSF to add questions for the purpose of research rather than consumer feedback.

Implications of segmentations for the program operators

As shown in this report, targeting health promotion messages by demographics such as gender (see pages 78-79) or age group (see pages 80-81) may not be practical or optimal, as differences may be small or there may be a potential to misclassify substantive proportions of the population. The limitations of demographics for understanding the health motivations, attitudes and behaviours of individuals have caused many health promotion practitioners to embrace "audience segmentation is one of the most important features of social marketing" (pg 8) (409). Segmentation is essential because "in developing effective health promotion strategies, it is more instructive to move beyond demographic variables (e.g., sex, age, and income) and to consider other variables that differentiate segments (e.g., unique needs, risk factors, propensity to change, and role as an influence on others)" (pg 5) (410). As well, as described by health promotion and social marketing consultant Heidi Keller in a TED talk, segmentation can guard against the natural bias of program developers to assume other people think the same way they do or value the same outcomes (411).

Commercial marketers often use four criteria to determine which segments to target (412):

- Measurability: whether a particular segment is large enough and has enough purchasing power or influence to produce a significant effect
- Accessibility: how easy it is to reach this segment
- Substantiality: whether the segment is large enough to make the investment in reaching them cost-effective
- Actionability: whether the segment is distinct enough to find and target
- Readiness to change: knowledge about the problem and readiness to change or to maintain the behaviour.

Using these criteria, it appears the k-means 2 solution is a viable segmentation for social marketing purposes as:

- The groups are of substantive size (measurability)
- The groups are easy to reach because they are all Internet health information seekers (accessibility); it is those outside these segments who may be difficult to reach because they are not online or are not health information-oriented
- Depending upon which segments are targeted, they should be large enough to justify investment in tailoring (substantiality)
- Within the HRA population and keeping in mind that segment boundaries are never distinct, the segments can be fairly easy to find by looking at patterns in age and modifiable risk factor and vascular disease burden (actionability)
- Two of the four segments appear already engaging in, or ready to engage in, health-enhancing behaviours (readiness to change); the other two may have the knowledge but either lack sufficient motivation or face barriers to engagement.

Thinking an etool is suitable for all of society is naïve. First, until the "digital divide" is breached, access to Internet-based resources will continue to vary by age, socioeconomic status and geography (57, 384, 413). Second, even with access, not all individuals are health information seekers (233). Operators of health etools could make better, more cost-efficient use of their resources if they used segmentation to learn more about their users, as well as their non-users. This information can then be used to: a) focus marketing efforts, b) tailor resources to meet the needs of user groups and increase messaging effectiveness, and c) develop alternative methods to reach those unlikely to be users. Making efficient use of funds is critical, given that health promotion and social marketing program typically have modest budgets (411).

Balancing program budgets and health promotion priorities can be difficult. University of South Florida social marketer R. Craig Lefebvre proposes that the choice of a target audience should be made by addressing three key questions (414):

- Who is at greatest risk?
- Who is most open to change?
- Who is it critical to reach in order to make your program successful?

In some respects, the HRA can be said to be performing adequately when it comes to reaching Canadians at risk who are Internet health information seekers. As shown, even though it was not effective in attracting those of lower SES or older age, 40% of users had one or more vascular diseases and 93.5% had one or more modifiable CVD risk factor, with the average being three risk factors. In other words, very few users can be said to be in a state of optimal cardiovascular health. Not only were HRA users of at least moderate CVD risk, but they appeared to be open to change, as demonstrated by the relatively high proportions (47% to 71%) in the Preparation or Action stages of change for various modifiable risk factors. These proportions would be strongly influenced by the self-selected, voluntary nature of open-access etools; nevertheless they provide rich opportunities for encouraging positive behaviour change among receptive participants.

When designing or operating health promotion programs that use mass media channels, there are challenges and hard choices sometimes need to be made. For example, Aboriginal Canadians (First Nations, Inuit and Métis) constitute a “high need” population for CVD prevention (354) but those living in remote or rural communities may not have broadband Internet access (413). Moreover, even if they have access, a health promotion etool would need to be designed and written to resonate with the three distinct Aboriginal cultures and the health challenges they face in their physical and socioeconomic environments. A substantive investment of resources would be required for a health etool with no guarantee of success, particularly in light of the health inequalities and social determinants of health for Aboriginal peoples (415).

In determining who they should target, organizations such as the HSF need to keep in mind that health risk assessments may not be tools for health behaviour change but rather “gateway interventions” (226) or “nudges” to help stimulate users to consider their lifestyle choices (228). Take, for example, the roughly third of HRA users who were categorized as being “less healthy” (the Younger and Less Healthy and Older and Less Healthy). Individuals in these groups appeared to be less committed to health-enhancing behaviours and more in need of behaviour change. But, given the limitations of secondary analysis of point prevalence data, our understanding of these people is incomplete. We cannot tell, for example, whether a less healthy lifestyle reflected a lack

of health conscientiousness or motivation or was caused by health-related physical limitations, social determinants, or other individual or community barriers (e.g., lack of access to healthier foods or recreational activities). For people in these groups, a "nudge" to consider behaviour change may be all an etool could be expected to accomplish.

The two healthier groups comprised almost two-thirds of HRA users and obviously constitute the "low hanging fruit" for the organization, in that these individuals displayed a higher level of health-enhancing behaviours suggesting greater health conscientiousness. With this sort of receptive audience, relatively low investments might have substantive impact on behaviour. However, although the easiest group to influence, would the return at the population level justify organizational investment? There is no one "correct" answer. The challenge is similar to that experienced when the population attributable fraction (PAF) of a disease is estimated for different risk factors. For example, smoking and physical inactivity have been estimated to each have a PAF for heart disease of 24% (416): the former because it greatly increases the risk but has a relatively low prevalence in the population and the latter because it has less impact upon heart disease risk but is more common. Organizations like the HSF need to consider what roles *need* (who is at greatest risk?), *receptivity* (who is most open to change?), and *reach* (who can we reach using this medium?) play in making program decisions (414).

Also important in organizational decision-making is answering Lefebvre's third question: who do you have to reach to consider your program successful (414)? The answer will vary according to the organization and its mandate. Although most health promotion programs hope to impact behaviour, attitudes and/or conditions to improve population health, the reality is that they can only affect clients or consumers – those who have direct interaction with their programs or services (417). Activities must not only be meaningful for clients/customers but have the potential of being sustainable and have value for the organization in terms of prestige, positioning, social/ethical benefits, or integration with its mission or vision (417).

In address Lefebvre's third question, the HSF needs to consider the critical audience for the HRA, both in isolation and as part of the organization's total inventory of health education/promotion activities. The answer will depend upon the function or functions the HSF envisions for the HRA and its sustainability. Is the etool considered an educational resource for online customers, a behaviour change tool for those unable or unwilling to access in-person or community services, a give-back for donors, or an incentive to attract new donors? Being either of the latter two does not take away the etool's potential for health promotion but would directly impact how success would be evaluated. It may be, for example, that other HSF programs services, such as its print

multicultural resources or its healthy public policy advocacy campaigns may have greater impact on Canadians with little or no Internet access. Etools may be still insufficient to reach all people in a large population with health needs.

Implications for other health promoters

It has been hypothesized that the Internet may be a powerful tool for health promotion and disease prevention (10, 12) because of its reach (368) and capacity for complex tailoring (12, 102, 418). These advantages have prompted many organizations to invest resources into the development and operation of online health etools. What is unclear is the extent to which organizations have recognized not only the strengths but also the weaknesses of the Internet when developing etools, of which one of the most important is dissemination biases by SES, gender and age (56, 57, 62, 74).

In their discussion of the open-access Heart Age calculator, Neufingerl *et al.* note: "The potential impact of Web-based health assessment tools on disease prevention is large" but to do so "they need to reach users with an elevated disease risk and provide accurate health assessments" (232). The authors admit the Heart Age calculator reaches "a large proportion of 'healthy' users" and a relatively small proportion of those at high risk (232). Likewise, previous analysis of the HSF population of users has noted that there may be under-representation of Canadians known to be at increased risk of CVD, such as males, seniors, smokers, and those of lower SES (230). However, given the still largely unexplored potential for the audiences of health etools to vary within the broad parameters of Internet health information seekers, it is difficult to make generalizations.

If *post hoc* analysis shows an organization that a specific etool is not reaching some of those at greatest need, what response would be appropriate? In the case of Heart Age, the authors have adopted a population-based philosophy and argued that "the audience reached by Heart Age *was* the intended audience – a group that is a good target for disease prevention" (italics in original) because it is currently largely CVD-free but has risk factors that elevate lifetime CVD risk (232). This approach is rooted in Geoffrey Rose's classic argument that a population-based strategy that focuses on societal norms may "shift the whole distribution of exposure in a favourable direction" and hence may have an impact on disease incidence "often larger than one would have expected" (419). Based on the distribution of risk, Rose argues that "a large number of people at a small risk may give rise to more cases of disease than the small number who are at a high risk" (419). In contrast, the high-risk strategy that stems from the traditional medical approach of treating individuals identified to be at risk may appear cost-effective when resources

are limited but this depends upon the existence of affordable and accurate screening, effective treatment, and appropriate uptake by those captured through screening (419).

In other words, one response to the fact that health etools tend to be used more by individuals who are health conscious may be to position it as part of a population-based approach to primary disease prevention. Given the digital divide, special efforts may be required to extend the reach beyond the typical Internet health information seeker to the population at large. Organizations may need to consider supplementary promotional strategies for the "harder to reach," such as advertising through ethnic media outlets (230) or mixing digital with off-line (e.g., print or in-person) resources or activities. Such approaches require a mix of complementary population and high-risk strategies, similar to what has been proposed in cardiology by Cooney *et al.* (420).

In the case of Heart Age (232) and the HRA (230), information about the biased utilization of the etool was released after the etool was already developed and launched through *post hoc* data analysis. As a result, if the decision is made to try and extend the reach of the etool (e.g., through specialized media promotion), these activities would be essentially reactive in nature. The advantage of a reactive response is that it can be evidence-based. The disadvantage, however, is that reacting may require additional investment of time, money or effort, such as developing new promotional campaigns, alternative delivery systems, or revising the etool look or language.

As previously discussed, in the case of Cancer Care Ontario's *My CancerIQ*™*,* market research helped to clarify which segments of the population would be most or least likely to utilize the site prior to site development or launch (396). The organization used this information to pro-actively develop strategies to extend the reach of the etool by partnering with public health and primary care providers (396) – those who are already in contact with some or all of the "harder to reach" segments. For example, some public health units have agreed to purchase tablets so they can give access to *My CancerIQ* to lower literacy or lower SES clients. A booth with tablets will also visit community hubs such as hospitals or shopping malls, providing visibility and access to the etool for those who might not have Internet access at home (396).

Not all organizations creating health etools have the funds needed to conduct pre-launch marketing research. However, all organizations could benefit by reviewing the literature on the characteristics of Internet health information seekers and taking it into account when developing their marketing and promotion plans. Understanding more about the type of people who utilize etools should inform decision-making about the proposed role or purpose of the etool (417) and potential or likely target audiences (412). Thinking through such issues should help avoid the common mistake of creating

information based more on the organization's standard messaging ("this is what we need to promote") than on the needs or interest of users (421). Accurate information about health etool users may make it possible for organizations to more effectively utilize digital health strategies among different types of Internet health information seekers. It may also help organizations understand which segments of the population may require different or alternative health promotion strategies.

Research strengths and limitations

Perhaps the greatest strength of this research is the amount of data available for analysis. In total, over 120,000 HRAs completed by individuals for themselves and with consent for research were available for analysis. The size of the data base made it possible to conduct sub-group analysis without having to worry about lacking power to detect significant differences.

At the same time, the database has several important limitations. The first is imposed by secondary analysis of an existing data base. The etool operator, the HSF, focuses almost exclusively on questions required for the generation of users' health risk reports. There was no opportunity to add questions to measure constructs or concepts that might be helpful in understanding the attitudes or motivation of users, such as health conscientiousness, self-efficacy, or health information orientation. Rather, such constructs had to be inferred from reported behaviours. Thus, the current research can be little more than exploratory research for hypothesis generation (331). Future research is needed that incorporates measurement of key constructs using questionnaires such as the Health Consciousness Scale (14), the Health Information Orientation Scale (35), Health Information Seeking Behaviours (90), and/or conscientiousness scale as part of the Big Five (422).

The second challenge concerns the validity of self-reported health data. Since the appropriateness of a user's health assessment report depends upon the honesty or accuracy of responses and no interviewer is involved, there is no apparent benefit or reason to "fudge" responses. Other research suggests self-reported health data may underestimate the proportion of individuals "at risk" or with risk factors (270), with the effect varying between different individuals, behaviours or conditions (271, 273, 275, 423). However, as discussed, there is also evidence that accuracy of web-based surveys may be better than those that are telephone-based (284), perhaps because of reduced socially acceptability (282) or acquiescence bias (283).

There was also no opportunity to conduct research to establish the effect of either the wording of questions or the order in which they were asked (259, 263). Thus, the extent to which questions were able to solicit knowledgeable and accurate information from the respondents is unknown (257). This is particularly true as in an open-access setting there can be no control over the context in which users respond, i.e., whether they are sincerely trying to assess their health or amusing themselves (87).

The lack of control over the context or motivation for completion of an online risk assessment is rooted in the open-access setting and the fact that users volunteer and self-select. Volunteer bias is a problem in much health and social research (363) and may be an unavoidable reality of open-access etools, as suggested by the considerable literature on the characteristics of Internet health information seekers (13, 48, 51, 57, 62, 65, 69, 79, 91, 95, 223, 348).

Finally, the sheer size of the HRA data base also created challenges, as even very small differences between groups were often statistically significant (297). Of the three challenges, this was the easiest to address. Effect sizes were used to separate small differences from those that were more substantive (i.e., medium- or large-sized).

Summary

Organizations that operate freely-available health etools have the potential to collect and analyze large amounts of information. To date, relatively little has been published on such etools, a gap that should be filled in order to expand knowledge and make possible the sharing of best and promising practices.

This analysis of a data base generated by the HRA operated by the HSF found etool users were neither a representative sample of the general population nor closely approximated by samples recruited for RCTs. Rather, although this self-selected population resembled previous research on health information seekers, it also appeared to be internally diverse. To dissolve the appearance of being a monolithic population, exploratory segmentation was conducted using three procedures and five sets of clustering variables. Several segmentations could be generated, the most robust and informative of which used k-means cluster analysis and two clustering variables that were moderately-to-strongly correlated (r=-.548) and incorporated readiness to change.

The four groups formed through the chosen segmentation differed by age (two younger and two older) and readiness to change, which may be an indicator of health

conscientiousness. The groups conformed to age-related trends observed in the HRA sample, such as the positive relationship between age and vascular diseases and the negative relationship between age and modifiable risk factors. One explanation is that trends reflect the observation that health conscientiousness, the ability and willingness to engage in health-enhancing behaviour, increases with age.

The segmentation provided new and potentially useful insights into the large "black box" of the HRA population that could be used by the etool operator in making decisions about what audience(s) to target, how to effectively tailor content to user needs, and whether alternative strategies may be needed to reach priority populations unlikely to be Internet users. In making these decisions, etool operators need to balance multiple, sometimes conflicting, factors of consumer need, receptivity, and corporate priorities. Given the complexity of this sort of decision-making, reliance on hunch or assumptions about who is likely to an etool is inadequate: organizations operating open-access health etools need analysis-based evidence.

Whether groups similar to those discovered in the HRA would emerge in other etool data bases is unknown. It is hoped this study might motivate organizations with similar data bases to undertake and share such analyses, thereby increasing our knowledge of who uses freely-available health etools, their needs and perhaps their motivators. Each analysis would undoubtedly have unique features, depending upon the type of etool, its target population and the data collected. To this end, the study emphasized procedures that are available in common software packages and discussed some of the analytic challenges other program operators may encounter and how they might be addressed. The overall message that organizations wishing to do similar analyses must understand is that generating a segmentation is not their primary challenge: software can easily pump out multiple solutions. The primary challenges are: 1) knowing the data base well enough that the choices of procedures, clustering variables and number of groups are informed, 2) embracing the idea that multiple segmentations may be generated; and 3) developing criteria by which to choose the optimal solution for the population being studied. Addressing these challenges will ensure that analyses are not exercises in "data dredging" but informed data mining for the purpose of knowledge discovery in data bases

References

1. Miniwatts Marketing Group. World Internet Usage and Population Statistics, March 31, 2011. Bogota: Miniwatts Marketing Group; 2012 [updated 2012; cited 2013 March 3]. Available from: http://www.internetworldstats.com/stats.htm.

2. Statistics Canada. Internet use by individuals, by selected frequency of use and age. Ottawa: Statistics Canada; 2010 [updated 2010/05/10; cited 2012 January 31]. Available from: http://www40.statcan.gc.ca/l01/cst01/comm32a-eng.htm.

3. Statistics Canada. Canadian Internet Use Survey, 2012. Ottawa: Statistics Canada. [updated 2013/11/26; cited 2013 November 8]. Available from: http://www.statcan.gc.ca/daily-quotidien/131126/dq131126d-eng.htm

4. Statistics Canada. Canadian Internet Use Survey, 2012. Ottawa: Statistics Canada 2013 Tuesday, November 26, 2013. Report No: 11-001-X.

5. U.S. Census Bureau. Table 1: Reported Internet Usage for Households, by Selected Housholder Characteristics: 2009. Current Population Survey (CPS) October 2009 (in thousands). Washington, D.C.: U.S. Census Bureau; 2010 [Updated February 2010; cited January 31, 2012]. Available from: http://www.census.gov/hhes/computer/publications/files/2009.html

6. Office for National Statistics. Internet Access - Housholds and Individuals - Tables, 2011. Newport, South Wales [Updated 2011 August 31; cited 2012 January 31]. Available from: http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-226727.

7. Gurak LJ, Hudson BL. E-Health: beyond internet searches. In: Murero M, Rice RE, editors. The Internet and Health Care, Theory, Research, and Practice. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2006. p. 29-48.

8. Annang L, Muilenburg JL, Strausser SM. Virtual worlds: taking health promotion to new levels. American Journal of Health Promotion. 2010;24(5):344-6.

9. Bennett GG, Glasgow RE. The delivery of public health interventions via the internet: actualizing their potential. Annual Review of Public Health. 2009;30:273-92.

10. Cassell M, Jackson C, Cheuvront B. Health communication on the Internet: an effective channel for health behavior change? Journal of Health Communication. 1998;3:71-9.

11. Cline RJW, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Education Research. 2001;16(6):671-92.

12. Stretcher V. Internet methods for delivering behavioral and health-related interventions (eHealth). Annual Review of Clinical Psychology. 2007;3:53-76.

13. Lambert SD, Loiselle CG. Health Information-Seeking Behavior. Qualitative Health Research. 2007;17(8):1006-19.

14. Hong H. Scale development of measuring health consciousness: re-conceptualization. International Public Relations Research Conference; Miaimi, Florida: 2009. Available from: http://www.instituteforpr.org/wp-content/uploads/ScaleDvlpmentMeasuring.pdf. Last accessed 2014 December 3

15. Gould SJ. Consumer Attitudes Toward Health and Health Care: A Differential Perspective. Journal of Consumer Affairs. 1988;22(1):96-118.

16. Kraft FB, Goodell PW. Identifying the health conscious consumer. Journal of Health Care Marketing. 1993;Fall:18-25.

17. Gould SJ. Health consciousness and health behavior: the application of a new health consciousness scale. Am J Prev Med. 1990;6(4):228-37.

18. Jayanti RK, Burns AC. The antecedents of preventive health care behavior: an empirical study. Journal of the Academy of Marketing Sciences. 1998;26(1):6-15.

19. Dutta-Bergman MJ. An Alternative Approach to Social Capital: Exploring the Linkage Between Health Consciousness and Community Participation. Health Communication. 2004;16(4):393-409.

20. Dissmore A, Gregersen E, Newberry A, Sonnenburg R. Self reported health consciousness levels: organic versus non-organic shoppers. Waukon, Iowa: Luther College: 2009. Available from:

http://www.slideshare.net/iowafoodandfitness/organic-vs-nonorganic-shoppers-presentation. Last accessed 2014 December 3

21. Wagner PJ, Curran P. Health beliefs and physician identified "worried well". Health Psychology. 1984;3(5):459-74.

22. Johansen SB, Næs T, Øyaas J, Hersleth M. Acceptance of calorie-reduced yoghurt: Effects of sensory characteristics and product information. Food Quality and Preference. 2010;21(1):13-21.

23. Lancaster KJ. Characteristics Influencing Daily Consumption of Fruits and Vegetables and Low-Fat Dairy Products in Older Adults with Hypertension. Journal of Nutrition For the Elderly. 2004;23(4):21-33.

24. Prasad A, Strijnev A, Zhang Q. What can grocery basket data tell us about health consciousness? International Journal of Research in Marketing. 2008;25(4):301-9.

25. Hoefkens C, Verbeke W, Van Camp J. European consumers' perceived importance of qualifying and disqualifying nutrients in food choices. Food Quality and Preference. 2011;22(6):550-8.

26. Park K, Choi KS, Kye SY, Park SH, Yoon NH, Park EC. Unwanted effects of risk notification for breast cancer regarding intention toward mammography utilization. Psycho-Oncology. 2010;19(8):823-9.

27. Merenstein DJ, Hu H, Robison E, Levine AM, Greenblatt R, Schwartz R, et al. Relationship Between Complementary/Alternative Treatment Use and Illicit Drug Use Among a Cohort of Women with, or at Risk for, HIV Infection. The Journal of Alternative and Complementary Medicine. 2010;16(9):989-93.

28. Bouwman LI, te Molder H, Koelen MM, van Woerkum CMJ. I eat healthfully but I am not a freak. Consumers' everyday life perspective on healthful eating. Appetite. 2009;53(3):390-8.

29. Newsom JT, McFarland BH, Kaplan MS, Huguet N, Zani B. The health consciousness myth: implications of the near independence of major health behaviors in the North American population. Social Science & Medicine. 2005;60(2):433-7.

30. Rakowski W, Julius M, Hickey T, Halter JB. Correlates of preventive health behavior in late life. Research on Aging. 1987;9(3):331-55.

31. Bloch PH. The wellness movement: imperatives for health care marketers. Journal of Health Care Marketing. 1984;4(1):9-16.

32. Roberts BW, Walton KE, Bogg T. Conscientiousness and health across the life course. Review of General Psychology. 2005;9(2):156-68.

33. Dutta-Bergman MJ. Developing a Profile of Consumer Intention to Seek Out Additional Information Beyond a Doctor: The Role of Communicative and Motivation Variables. Health Communication. 2005;17(1):1-16.

34. Dutta-Bergman MJ. Media use theory and internet use for health care. In: Murero M, Rice RE, editors. The Internet and Health Care, Theory, Research, and Practice. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2006. p. 83-103.

35. DuBenske LL, Burke BE, Hawkins RP, Gustafson DH. Psychometric Evaluation of the Health Information Orientation Scale A Brief Measure for Assessing Health Information Engagement and Apprehension. Journal of Health Psychology. 2009;14(6):721-30.

36. Sweeny K, McInyk D, Miller W, Shepperd JA. Information avoidance: who, what, when, and why. Review of General Psychology. 2010;14(4):340-53.

37. Miller S. Monitoring and blunting: validation of a questionnaire to assess styles of information. Journal of Personality and Social Psychology. 1987;52(2):345-53.

38. Wolff LS, Massett HA, Maibach EW, Weber D, Hassmiller S, Mockenhaupt RE. Validating a health consumer segmentation model: behavioral and attitudinal differences in disease prevention-related practices. Journal of Health Communication. 2010;15:167-88.

39. Jackson JJ, Wood D, Bogg T, Walton K, Harms P, Roberts B. What do conscientious people do? Development and validiation of the Behavioral Indicators of Conscientiousness (BIC). Journal of Research in Personality. 2010;44(4):501-11.

40. Baumeister RF, Gailliot M, Dewall CN, Oaten H. Self-regulation and personality: how interventions increase regulatory success and how depletion moderates the effects of traits on behavior. Journal of Personality. 2006;74(6):1773-802.

41. Hofman W, Baumeister RF, Forster G, Vohs KD. Everday temptations: an experience sampling study of desire, conflict, and self-control. Journal of Personality and Social Psychology. 2012;102(6):1318-35.

42. Bogg T, Roberts BW. Conscientiousness and health-related behaviors: a meta-analysis of the leading behavioral contributors to mortality. Psychological Bulletin. 2004;130(6):887-919.

43. Lodi-Smith J, Jackson J, Bogg T, Walton K, Wood D, Harms P, et al. Mechanisms of health: Education and health-related behaviours partially mediate the relationship between conscientiousness and self-reported physical health. Psychology & Health. 2010;25(3):305-19.

44. Kern ML, Friedman HS. Conscientiousness, career success, and longevity: a lifespan analysis. Annals of Behavioral Medicine. 2009;37(2):154-63.

45. Jackson JJ, Bogg T, Walton KE, Wood D, Harms PD, Lodi-Smith J, et al. Not all conscientiousness scales change alike: a multimethod, multisample study of age differences in the facets of conscientiousness. Journal of Personality and Social Psychology. 2009;96(2):446-59.

46. Takashi Y, Edmonds GW, Jackson JJ, Roberts BW. Longitudinal correlated changes in conscientiousness, preventative health-related behaviors, and self-perceived physical health. Journal of Personality. 2013;81(4):417-27.

47. Hill PL, Roberts B. The role of adherence in the relationship between conscientiousness and perceived health. Health Psychology. 2011;30(6):797-804.

48. Bundorf M, Wagner T, Singer S, Baker L. Who searches the internet for health information? Health Services Research. 2006;41(3):819-36.

49. Baker L, Wagner TH, Singer S, Bundorf MK. Use of the Internet and E-mail for Health Care Information. JAMA: The Journal of the American Medical Association. 2003;289(18):2400-6.

50. Beaudoin CE, Hong T. Health information seeking, diet and physical activity: An empirical assessment by medium and critical demographics. International Journal of Medical Informatics. 2011;80(8):586-95.

51. Fox S. The Social Life of Health Information, 2011. Washington, D.C.: Pew Internet and American Life Project, 2011. [updated 2011 May 12; cited 2013

November 11]. Available from: http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/

52. Taylor H. "Cyberchondriacs" on the Rise? Those who go online for healthcare information continues to increase. The Harris Poll, 2010. [Updated 2010 August 4; cited 2013 Nov 3]. Available from: http://www.harrisinteractive.com/NewsRoom/HarrisPolls/tabid/447/mid/1508/articleId/448/ctl/ReadCustom%20Default/Default.aspx

53. Statistics Canada. Table 2. Online activities of home Internet users. Statistics Canada; Ottawa; 2010 [Updated 2013 October 28; cited 2012 January 31]. Available from: http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/comm29a-eng.htm

54. Statistics Canada. Individual Internet use and e-commerce, 2012. Ottawa: Statistics Canada, 2013 [Updated 2013 October 28; cited 2013 November 11]. Available from: http://www.statcan.gc.ca/daily-quotidien/131028/dq131028a-eng.htm

55. Office for National Statistics. Internet Access 2008. Households and Individuals. Newport, South Wales: 2008 [Updated 2010 September 9; cited 2013 November 11]. Available from: http://www.ons.gov.uk/ons/index.html

56. Schneider F, van Osch LADM, Kremers SPJ, Schulz DN, Van Adrichem MJG, de Vires H. Optimizing diffusion of an online computer tailored lifetyle program: a study protocol. BMC Public Health. 2011;11:480.

57. Choi MG. The digital divide among low-income homebound older adults: internet use patterns, ehealth literacy, and attitudes toward computer/internet information. Journal of Medical Internet Research. 2013;15(5):e93.

58. Bass SB, Ruzek SB, Gordon TF, Fleisher L, McKeown-Conn N, Moore D. Relationship of Internet Health Information Use With Patient Behavior and Self-Efficacy: Experiences of Newly Diagnosed Cancer Patients Who Contact the National Cancer Institute's Cancer Information Service. Journal of Health Communication. 2006;11(2):219-36.

59. Beaudin JS, Intille SS, Morris ME. To track or not to track: user reactions to concepts in longitudinal health monitoring. Journal of Medical Internet Research. 2006;8(4):e29.

60. Callas PW, Solomon LJ, Hughes JR, Livingston AE. The influence of response mode on study results: offering cigarette smokers a choice of postal or online completion of a survey. Journal of Medical Internet Research. 2010;12(4):e48.

61. Chisolm DJ. Does online health information seeking act like a health behavior?: a test of the behavioral model. Telemedicine and e-Health. 2010;16(2):154-60.

62. Cotten SR, Gupta SS. Characteristics of online and offline health information seekers and factors that discriminate between them. Social Science & Medicine. 2004;59(9):1795-806.

63. de Boer MJ, Versteegen GJ, van Wijhe M. Patients' use of the internet for pain-related medical information. Patient Education and Counseling. 2007;68:86-97.

64. Hou J, Shim M. The Role of Provider-Patient Communication and Trust in Online Sources in Internet Use for Health-Related Activities. Journal of Health Communication: International Perspectives. 2010;15(S3):186-99.

65. Atkinson NL, Saperstein SL, Pleis J. Using the Internet for health-related activities: findings from a national probability sample. Journal of Medical Internet Research. 2009;11(1).

66. Cohall AT, Nye A, Moon-Howard J, Kukafka R, Dye B, Vaught RD, et al. Computer use, internet access, and online health searching among Harlem adults. American Journal of Health Promotion. 2011;25(5):325-33.

67. Kim K, Kwon N. Profile of e-Patients: Analysis of Their Cancer Information-Seeking From a National Survey. Journal of Health Communication. 2010;15(7):712-33.

68. Kind T, Huang ZJ, Farr D, Pomerantz KL. Internet and computer access and use for health information in an underserved community. Ambulatory Pediatrics. 2005;5(2):117-21.

69. Koch-Weser S, Bradshaw YS, Gualtieri L, Gallagher SS. The Internet as a Health Information Source: Findings from the 2007 Health Information National Trends Survey and Implications for Health Communication. Journal of Health Communication: International Perspectives. 2010;15(S3):279-93.

70. Powell JA, Darvell M, Gray JAM. The doctor, the patient and the world-wide web: how the internet is changing healthcare. Journal of the Royal Society of Medicine. 2003;96(2):74-6.

71. Wong B, Yung B, Wong A, Chow C, Abramson B. Increasing internet use among cardiovascular patients: new opportunities for heart health promotion. Canadian Journal of Cardiology. 2005;21(4):349-54.

72. Zulman D, Kirch MZ, K, An L. Trust in the internet as a health resource among older adults: analysis of data from a nationally representative survey. Journal of Medical Internet Research. 2011;13(1):e19.

73. Flynn KE, Smith MA, Freese J. When Do Older Adults Turn to the Internet for Health Information? Findings from the Wisconsin Longitudinal Study. Journal of General Internal Medicine. 2006;21(12):1295-301.

74. Renahy E, Parizot I, Chauvin P. Health information seeking on the Internet: a double divide? Results from a representative survey in the Paris metroplitan area, France, 2005-2006. BMC Public Health. 2008;8:69.

75. Dickerson S, Reinhart AM, Feeley TH, Bidani R, Rich E, Garc VK, et al. Patient internet use for health information at three urban primary care clinics. Journal of the American Medical Informatics Association. 2004;11(6):499-504.

76. Kalichman SC, Weinhardt L, Benotsch E, DiFonzo K, Luke W, Austin J. Internet access and internet use for health information among people living with HIV-AIDS. Patient Education and Counseling. 2002;46:109-16.

77. Wangberg SC, Andreassen HK, Prokosch H-U, Santana SMV, Sørensen T, Chronaki CE. Relations between Internet use, socio-economic status (SES), social support and subjective health. Health Promotion International. 2008;23(1):70-7.

78. Houston TK, Allison JJ. Users of internet health information: differences by health status. Journal of Medical Internet Research. 2002;4(2):e7.

79. Dutta M, Feng H. Health Orientation and Disease State as Predictors of Online Health Support Group Use. Health Communication. 2007;22(2):181-9.

80. Dutta-Bergman M. Health attitudes, health cognitions, and health behaviors among internet health information seekers: population-based survey. Journal of Medical Internet Research. 2004;6(2):e15.

81. Pandey SK, Hart JJ, Tiwary S. Women's health and the internet: understanding emerging trends and implications. Social Science & Medicine. 2003;56(1):179-91.

82. Moorman C, Matulich E. A Model of Consumers' Preventive Health Behaviors: The Role of Health Motivation and Health Ability. Journal of Consumer Research. 1993;20(2):208-28.

83. Goldner M. How health status impacts the types of information consumers seek online. Information, Communication & Society. 2006;9(6):693-713.

84. Dutta-Bergman MJ. Primary Sources of Health Information: Comparisons in the Domain of Health Attitudes, Health Cognitions, and Health Behaviors. Health Communication. 2004;16(3):273-88.

85. Cho J, Park D, Lee H. Cognitive factors of using health apps: systematic analysis of relationships among health consciousness, health information orientation, eHealth literacy, and health app use efficacy. Journal of Medical Internet Research. 2014;16(5):e125.

86. Kahlor L. PRISM: a Plann Risk Information Seeking Model. Health Communication. 2010;25:345-56.

87. Boot CRL, Meijman FJ. The public and the internet: multifaceted drives for seeking health information. Health Informatics Journal. 2010;16(2):145-56.

88. Poirier J, Cobb NK. Social influence as a driver of engagement in a web-based health intervention. Journal of Medical Internet Research. 2012;14(1):e36.

89. Choi N, DiNitto DM. Internet use among older adults: association with health needs, psychological capital, and social capital. Journal of Medical Internet Research. 2013;15(5):e97.

90. Weaver JB, Mays D, Sargent Weaver S, Hopkins GL, Eroglu D, Bernhardt JM. Health information-seeking behaviors, health indicators, and health risks. American Journal of Public Health. 2010;100(8):1520-25.

91. Goldner M. Using the Internet and Email for Health Purposes: The Impact of Health Status. Social Science Quarterly. 2006;87(3):690-710.

92. Ayers SL, Kronenfeld JJ. Chronic illness and health-seeking information on the Internet. Health:. 2007;11(3):327-47.

93. Eastin MS, Guinsler NM. Worried and Wired: Effects of Health Anxiety on Information-Seeking and Health Care Utilization Behaviors. CyberPsychology & Behavior. 2006;9(4):494-8.

94. Lee SY, Hwang H, Hawkins R, Pingree S. Interplay of negative emotion and health self-efficacy on the use of health information and its outcomes. Communication Research. 2008;35(3):358-81.

95. Lemire M, Paré G, Sicotte C, Harvey C. Determinants of Internet use as a preferred source of information on personal health. International Journal of Medical Informatics. 2008;77(11):723-34.

96. Wilkins ST, Navarro FH. Has the web really empowered health care consumers? Marketing Health Services. 2001;Fall:5-9.

97. Fox S, Duggan M. Health Online 2013. Washington, DC: Pew Research Center, 2013. Available from: http://www.pewinternet.org/2013/01/15/health-online-2013/ Last accessed 2014 December 13

98. Evers KE. eHealth promotion: the use of the Internet for health promotion. The Art of Health Promotion. 2006;March/April:1-7.

99. Quorus Consulting Group. 2012 Cell Phone Consumer Attitudes Study. Ottawa: Canadian Wireless Telecommunications Association, 2012 April 23, 2012. Available from: http://cwta.ca/wordpress/wp-content/uploads/2011/08/CWTA-2012ConsumerAttitudes.pdf Last accessed 2014 December 3.

100. Lefebvre RC, Tada Y, Hilfiker SW, Baur C. The assessment of user engagement with eHealth content: The eHealth Engagement Scale. Journal of Computer-Mediated Communications. 2010;15:666-81.

101. Baek TH, Yu H. Online health promotion strategies and appeals in the USA and South Korea: a content analysis of weight-loss websites. Asian Journal of Communication. 2009;19(1):18 - 38.

102. Kreuter MW, Farrell D, Olevitch L. Tailoring Health Messages, Customizing Communication with Computer Technology. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2000.

103. Kunst H, Groot D, Latthe PM, Latthe M, Khan KS. Accuracy of information on apparently credible websites: survey of five common health topics. BMJ. 2002;324(7337):581-2.

104. Evers K, Prochaska J, Prochaska J, Driskell M, Cumins C, WF V. Strengths and weaknesses of health behavior change programs on the Internet. Journal of Health Psychology. 2003;8:63-70.

105. Webb TL, Joseph J, Yadley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. Journal of Medical Internet Research. 2010;12(1):e4.

106. Michie S, Johnston M, Francis JJ, Hardeman W, Eccles MP. From theory to intervention: mapping theoretically dervied behavioural determinants to behaviour change techniques. Applied Psychology: An International Review. 2008;57(4):660-80.

107. Sillence E, Briggs P, Harris P, Fishwick L. A framework for understanding trust factors in web-based health advice. International Journal of Human-Computer Studies. 2006;64:697-713.

108. Harris PR, Sillence E, Briggs P. The effect of credibility-related design cues on responses to a web-based message about the breast cancer risks from alcohol: randomized controlled trial. Journal of Medical Internet Research. 2009;11(3):e37.

109. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Social Sciences & Medicine. 2007;64:1853-62.

110. Cugelman B, Thelwall M, Dawes P. Online interventions for social marketing health behaviour change campaigns: a meta-analysis of psychological architectures and adherence factors. Journal of Medical Internet Research. 2010;13(1):e17.

111. Danaher B, McKay H, Seeley J. The information architecture of behaviour change websites. Journal of Medical Internet Research. 2005;7(2):e12.

112. Alexander GL, Divine GW, Couper MP, McClure JB, Stopponi MA, Fortman KK, et al. Effect of incentives and mailing features on online health program enrollment. American Journal of Preventive Medicine. 2008;34(5):382-88.

113. Bull S, Vallejos D, Levine D, Ortiz C. Improving recruitment and retention for an online randomized controlled trial: experience from the Youthnet study. AIDS Care. 2008;20(8):887-93.

114. Eysenbach G. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ. 2002;324(7337):573-7.

115. Kelly L, Jenkinson C, Ziebland S. Measuring the effects of online health information for patients: Item generation for an e-health impact questionnaire. Patient Education and Counseling. 2013;93:433-8.

116. Danaher BG, Seeley JR. Methodological issues in research on web-based interventions. Annals of Behavioral Medicine. 2009;38(1):23-39.

117. Civljak M, Sheilkh A, Stead LR, Car J. Internet-based interventions for smoking cessation (Review). Cochrane Library of Systematic Reviews. 2010;9:Art. No: CD007078.

118. Shahab L, McEwen A. Online support for smoking cessation: a systematic review of the literature. Addiction. 2009;104(11):1792-804.

119. Bailey JV, Murray E, Rait G, Mercer CH, Morris RW, Peacock R, et al. Interactive computer-based interventions for sexual health promtion. Cochrane Database of Systematic Reviews. 2010;9:CD006483.

120. Samoocham D, Bruinvels DJ, Elbers NA, Anema JR, van der Beek AJ. Effectiveness of web-based interventions on patient empowerment: a systematic review and meta-analysis. Journal of Medical Internet Research. 2010;12(2):e34.

121. Cook R, Billings D, Hersch R, BACK A, Hendrickson A. A field test of a web-based workplace health promotion program to improve dietary practices, reduce stress, and increase physical activity: randomized controlled trial. Journal of Medical Internet Research. 2007;9(2).

122. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professionas: a meta-analysis. JAMA. 2008;300:1181-96.

123. Wantland D, Portillo C, Holzemer W, Slaughter R, McGhee E. The effectiveness of web-based vs. non-web-based interventions: a meta-analysis of behavioral change outcomes. Journal of Medical Internet Research. 2004;6(4):e40.

124. Carey KB, Scott-Sheldon LA, Elliott JC, Bolles JR, Carey MP. Computer-delievered interventions to reduce college student drinking: a meta-analysis. Addiction. 2009;104:1807-19.

125. Neve M, Morgan PJ, Jones PR, Collins C. Effectiveness of web-based interventions in achieving weight loss and weight loss maintenance in overweight and obese adults: a systematic review with met-analysis. Obesity Review. 2010;11:306-21.

126. Reed VA, Schifferdecker KE, Rezaee ME, O'Connor S, Larson RJ. The effect of computers for weight loss: a systematic review and meta-analysis of randomized trials. Journal of General Internal Medicine. 2012;27(1):99-108.

127. Noar SM, Blac HG, Pierce LB. Efficacy of computer technology-based HIV prevention interventions: a meta-analysis. AIDS 2009;23:107-15.

128. Spek V, Cuijpers P, Nyklicek I, Riper H, Keyzer J, Pop V. Internet-based cognitive behavior therapy for symptoms of depression and anxiety: a meta-analysis. Psychological Medicine. 2007;38:310-28.

129. Murray E, Burns J, Tai S, R. L, Nazareth I. Interactive health communication applications for people with chronic disease (review). The Cochrane Library. 2005(4):Art. No.: CD004274.

130. Stinson J, Wilson R, Gill N, Yamada J, J. H. A systematic review of internet-based self-management interventions for youth with health conditions. Journal of Pediatric Psychology. 2009;34:495-510.

131. Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M. Why are health care interventions delivered over the Internet? A systematic review of the published literature. Journal of Medical Internet Research. 2006;8(2).

132. Griffiths KM, Farrer L, Christensen H. The efficacy of internet interventions for depression and anxiety disorders: review of randomized controlled trials. Medical Journal of Australia. 2010;192(11 Suppl):S4-S11.

133. Christensen H, Griffiths J, Farrer L. Adherence in internet interventions for anxiety and depression: systematic review. Journal of Medical Internet Research. 2009;11(2):e13.

134. Calear AL, Christensen H. Review of internet-based prevention and treatment programs for anxiety and depression in children and adolescents. Medical Journal of Australia. 2010;192(11 Suppl.):S12-S4.

135. Van't Hof E, Cuijpers P, Stein DJ. Self-help and internet-guided interventions in depression and anxiety disorders: a systematic review of meta-analyses. CNS Spectr. 2009;14(2 Suppl 3):34-40.

136. Newton MS, Cilska D. Internet-based innovations for the prevention of eating disorders: a systematic review. Eating Disorders. 2006;14:365-84.

137. Tait RJ, Christensen H. Internet-based interventions for young people with problematic substance use: a systematic review. Medical Journal of Australia. 2010;192(11 Supp):S15-S21.

138. Enwald HPK, Huotari M-LA. Preventing the obesity epidemic by second generation tailored health communication: an interdisciplinary review. Journal of Medical Internet Research. 2010;12(2):e24.

139. van den Berg MH SJ, Vliet Vlieland TPM. Internet-based physical activity interventions: a systematic review of the literature. Journal of Medical Internet Research. 2007;9(3):e26.

140. An JY, Hayman LL, Park YS, Dusaj TK, Ayres CG. Web-based weight management programs for children and adolescents: a systematic review of randomized controlled trials. ANS Advances in Nursing Science. 2009;32:222-40.

141. Norman GJ, Zabinski MF, Adams MA, Rosenberg DE, Yaroch AL, Atienza AA. A Review of eHealth Interventions for Physical Activity and Dietary Behavior Change. American Journal of Preventive Medicine. 2007;33(4):336-45.e16.

142. Sanchez MA, Rabin BA, Gaglio B, Henton M, Elzarrad MK, Purcell P, et al. A systematic review of eHealth cancer prevention and control interventions: new technology, same methods and designs? Translational Behavioral Medicine. 2013;3:392-401.

143. Portnoy DB, Schott-Sheldon LAJ, Johnson BT, Carey MP. Computer-delivered interventions for health promotion and behavioral risk reduction: A meta-analysis of 75 randomized controlled trials, 1988-2007. Preventive Medicine. 2008;47:3-16.

144. Manzoni GM, Pagnini F, Corti S, Molinari E, Castelnuovo G. Internet-based behavioral interventions for obesity: an updated systematic review. Clinical Practice & Epidemiology in Mental Health. 2011;7:19-28.

145. Arem H, Irwin M. A review of web-based weight loss interventions in adults. Obesity Review. 2011;12(5):e236-e43.

146. Krukowski RA, West DS, Harvey-Berino J. Recent advances in internet-delivered, evidence-based weight control programs for adults. Journal of Diabetes Science and Technology. 2009;3(1):184-89.

147. Neville LM, O'Hara B, Milat AJ. Computer-tailored physical activity behavior change interventions targeting adults: a systematic review. International Journal of Behavioral Nutrition and Physical Activity. 2009;6:30.

148. Weinstein PK. A review of weight loss programs devliered via the internet. Journal of Cardiovascular Nursing. 2006;21(4):251-8.

149. Bennett GG, Herring SJ, Puleo E, Stein E, Emmons KM, Gillman MW. Web-based weight loss in primary care: a randomized controlled trial. Obesity. 2010;18:308-13.

150. Blanson Henkemanns OA, van der Boog PJM, Lindenberg J, van der Mast CAPG, Neerincx MA, Zwetsloot-Schonk BJHM. An online lifestyle diary with a persausive computer assistant providing feedback on self-management. Technology and Health Care. 2009;17:253-67.

151. Booth AO, Nowson CA, Matters H. Evaluation of an interactive, Internet-based weight loss program: a pilot study. Health Education Research. 2008:cyn007.

152. Cussler EC, Teixeira PJ, Going SB, Houtkooper LB, Metcalfe LL, Blew RM, et al. Maintenance of Weight Loss in Overweight Middle-aged Women Through the Internet. Obesity. 2008;16(5):1052-60.

153. Glasgow R, Nelson C, Kearney K, Reid R, Ritzwoller D, Strecher V, et al. Reach, engagement and retention in an Internet-based weight loss program in a multi-site randomized controlled trial. Journal of Medical Internet Research. 2007;9(2).

154. Gold BC, Burke S, Pintauro S, Buzzell P, Harvey-Berino J. Weight loss on the web: a pilot study comparing a structured behavioral intervention to a commercial program. Obesity. 2007;15(1):155-64.

155. Hageman P, Pullen C, Hertzog M, Boeckner LS, Walker SN. Web-based interventions for weight loss and weight maintenance among rural midlife and

older women: protocol for a randomized controlled trial. BMC Public Health. 2011;11:521.

156. Harvey-Berino J, Pintauro S, Buzzell P, DiGiulio M, Casey Gold B, Moldovan C, et al. Does using the Internet facilitate the maintenance of weight loss? International Journal of Obesity. 2002;26(9):1254-60.

157. Harvey-Berino J, Pintauro S, Buzzell P, Gold EC. Effect of Internet Support on the Long-Term Maintenance of Weight Loss. Obesity. 2004;12(2):320-9.

158. Harvey-Berino J, Pintauro SJ, Gold EC. The Feasibility of Using Internet Support for the Maintenance of Weight Loss. Behavior Modification. 2002;26(1):103-16.

159. Hunter CM, Peterson AL, Alvarez LM, Poston WC, Brundige AR, Haddock K, et al. Weight management using the Internet - A randomized controlled trial. American Journal of Preventive Medicine. 2008;34(2):119-26.

160. Kelders S, van Gemert-Pijnen J, Wekman A, Seydel E. Usage and effect of a web-based intervention for the prevention of overweight; a RCT. Studies in Health Technology and Informatics. 2010 160 (Pt 1):28-32.

161. McConnon A, Kirk S, Cockroft J, Harvey E, Greenwood D, Thomas J, et al. The Internet for weight control in an obese sample: results of a randomised controlled trial. BMC Health Services Research. 2007;7(1):206.

162. Morgan PJ, Lubans DR, Collins CE, Warren JM, Callister R. The SHED-IT randomized controlled trial: evaluation of an internet-based weight-loss program for men. Obesity. 2009;17(11):2025-32.

163. Patrick K, Raab F, Adams M, Dillong L, Zabinski M, Rock C, et al. A text message-based intervention for weight loss: randomized controlled trial. Journal of Medical Internet Research. 2009;11(1):e1.

164. Polzien KM, Jakicic JM, Tate DF, Otto AD. The Efficacy of a Technology-based System in a Short-term Behavioral Weight Loss Intervention. Obesity. 2007;15(4):825-30.

165. Pullen CH, Hageman P, Boeckner LS, Walker S, Oberdorfer MK. Feasibility of internet-delivered weight loss interventions among rural women ages 50-69. Journal of Geriatric Physical Therapy. 2008;31(3):105-12.

166. Rothert K, Strecher VJ, Doyle LA, Caplan WM, Joyce JS, Jimison HB, et al. Web-based Weight Management Programs in an Integrated Health Care Setting: A Randomized, Controlled Trial. Obesity. 2006;14(2):266-72.

167. Franko DL, Cousineau TM, Trant M, Green TC, Rancourt D, Thompson D, et al. Motivation, self-efficacy, physical activity and nutrition in college students: Randomized controlled trial of an internet-based education program. Preventive Medicine. 2008;47(4):369-77.

168. Smeets T, Kremer S, de Vries H, Brug J. Effects of tailored feedback on multiple health behaviours. Annals of Behavioral Medicine. 2007;33(2):117-23.

169. Tate DF, Wing RR, Winett RA. Using Internet Technology to Deliver a Behavioral Weight Loss Program. JAMA. 2001;285(9):1172-7.

170. Tate DF, Jackvony EH, Wing RR. Effects of Internet Behavioral Counseling on Weight Loss in Adults at Risk for Type 2 Diabetes: A Randomized Trial. JAMA. 2003;289(14):1833-6.

171. Turner-McGrievy GM, Campbell MK, Tate DF, Truesdale KP, Bowling JM. Pounds Off Digitally Study. A randomized podcasting weight-loss intervention. American Journal of Preventive Medicine. 2009;2009(37):4.

172. van Wier M, Ariens G, Dekkers JC, Hendriksen I, Smid T, van Mechelen W. Phone and e-mail counselling are effective for weight management in an overweight working population: a randomized controlled trial. BMC Public Health. 2009;9(1):6.

173. Webber K, Tate D, Bowling J. A randomized comparison of two motivationally enhanced Internet behavioral weight loss programs. Behaviour Research and Therapy. 2008;46:1090-5.

174. Webber K, Tate DF, Ward D, Bowling JM. Relationships among motivation, adherence, and weight loss in a 16-week Internet behavioral weight loss intervention. Annals of Behavioral Medicine. 2008;35:S168-S.

175. Webber KH, Tate DF, Ward DS, Bowling JM. Motivation and Its Relationship to Adherence to Self-monitoring and Weight Loss in a 16-week Internet Behavioral Weight Loss Intervention. Journal of Nutrition Education and Behavior. 2010;42(3):161-7.

176. White MA, Martin PD, Newton RL, Walden HM, York-Crowe EE, Gordon ST, et al. Mediators of Weight Loss in a Family-Based Intervention Presented over the Internet. Obesity. 2004;12(7):1050-9.

177. Womble LG, Wadden TA, McGuckin BG, Sargent SL, Rothman RA, Krauthamer-Ewing ES. A Randomized Controlled Trial of a Commercial Internet Weight Loss Program Obesity. 2004;12(6):1011-8.

178. McTigue KM, Conroy MB, Hess R, Bryce CL, Fiorillo AB, Fischer GS, et al. Using the internet to translate an evidence-based lifestyle intervention into practice. Telemedicine and e-Health. 2009;15(9):851-8.

179. Winett RA, Tate DF, Anderson ES, Wojcik JR, Winett SG. Long-term weight gain prevention: A theoretically based Internet approach. Prev Med. 2005;41(2):629-41.

180. Oenema A, Dijstra A, de Vires H. Efficacy and use of an internet-delivered computer-tailored lifestyle intervention, targeting saturated fat intake, physical activity and smoking cessation: a randomized controlled trial. Annals of Behavioral Medicine. 2008;35:125-35.

181. Neve M, Morgan PJ, Collins CE. Weight change in a commercial web-based weight loss program and its association with website use: cohort study. Journal of Medical Internet Research. 2011;13(4):e83.

182. Carter-Edwards L, Bastian LA, Schultz M, Amamo MA, Ostbye T. An internet-based weight loss intervention initiated by a newspaper. Preventing Chronic Disease. 2009;6(3).

183. Haugen HA, Tran ZV, Wyatt HR, Barry MJ, Hill JO. Using Telehealth to Increase Participation in Weight Maintenance Programs. Obesity. 2007;15(12):3067-77.

184. Johnson F, Wardle J. The association between weight loss and engagement with a web-based food and exercise diary in a commercial weight loss programme: a retrospective analysis. Journal of Behavioral Nutrition and Physical Activity. 2011;8:83.

185. Jonasson J, Linne Y, Neovius M, Rossner S. An Internet-based weight loss programme - a feasibility study with preliminary results from 4209 completers. Scandinavian Journal of Public Health. 2009;37(1):75-82.

186. Kelders SM, Van Gemert-Pijnen J, Werkman A, Seydel ER. Evaluation of a web-based lifestyle coach designed to maintain a healthy bodyweight. Journal of Telemedicine and Telecare. 2010;16:3-7.

187. Krukowski R, Harvey-Berino J, Ashikaga T, Thomas C, Micco N. Internet-based weight control: the relationship between web features and weight loss. Telemedicine and e-Health. 2007;14(8):775-82.

188. Bensley R, Brusk J, Rivas J. Key principles in internet-based weight management systems. American Journal of Health Behavior. 2010;34(2):206-13.

189. Jung T, McClung S, Youn H, Chang T-S. Losing weight on the Web? A content analysis of dieting-related Web sites. He@lth Information on the Internet. 2007;59(59):3-6.

190. Saperstein SL, Atkinson NL, Gold RS. The impact of Internet use for weight loss. Obesity Review. 2007;8(5):459-65.

191. Binks M, Van Mierio T. Utilization patterns and user characteristics of an ad libitum internet weight loss program. Journal of Medical Internet Research. 2010;12(1):e9.

192. Neve M, Collins C, Morgan P. Dropout, nonusage attrition, and pretreatment predictors of nonusage attrition in a commercial web-based wieght loss program. Journal of Medical Internet Research. 2010;12(4):e69.

193. You W, Almeida FA, Zoellner JM, Hill JL, Pinard CA, Allen KC, et al. Who participates in internet-based worksite weight loss programs? BMC Public Health. 2011;11:709.

194. Maitland J, Chalmers M, editors. Finding a balance: social support v. privacy during weight-management. 26th Annual CHI Conference on Human Factors in Computing Systems; 2008 April 5 - April 10, 2008; Florence, Italy. ACM Ditial Library: Association for Computing Machinery (ACM); 2008.

195. Hwang KO, Ottenbacher AJ, Green AP, Cannon-Diehl MR, Richardson O, Bernstam EV, et al. Social support in an internet weight loss community. International Journal of Medical Informatics. 2010;79:5-13.

196. McTigue KM, Bhargava T, Bryce DL, Conroy MB, Fischer GS, Hess R, et al. Patient perspectives on the integration of an intensive online behavioral

weight loss intervention into primary care. Patient Education and Counseling. 2011;83:261-4.

197. Plotnikoff RC, Pickering MA, McCargar LJ, Loucaides CA, Hugo K. Six-month follow-up and participant use and satisfaction of an electronic mail intervention promoting physical activity and nutrition. American Journal of Health Promotion. 2010;24(4):255-9.

198. Ware LJ, Hurling R, Bataveljic O, Fairley BW, Hurst TL, Murray P, et al. Rates and Determinants of Uptake and Use of an Internet Physical Activity and Weight Management Program in Office and Manufacturing Work Sites in England: Cohort Study. Journal of Medical Internet Research. 2008;10(4):17.

199. Koo M, Skinner H. Challenges of internet recruitment: a case study with disapponting results. Journal of Medical Internet Research. 2005;7(1):e6.

200. Stevens V, Funk K, Brantley P, Erlinger T, Myers V, Champagne C, et al. Design and implementation of an interactive website to support long-term maintenance of weight loss. Journal of Medical Internet Research. 2008;10(1):e1.

201. Winnet R, Anderson E, Wojcik J, Winnett S, Bowden T. Guide to Health: Nutrition and physical activity outcomes of a group-randomized trial of an internet-based intervention in churches. Annals of Behavioral Medicine. 2007;33(3):251-61.

202. Blenkinsopp J. Obesity. He@lth Information on the Internet. 2007;55(1):6.

203. Moore TJ, Alsabeeh N, Apovian CM, Murphy MC, Coffman GA, Cullum-Dugan D, et al. Weight, Blood Pressure, and Dietary Benefits After 12 Months of a Web-based Nutrition Education Program (DASH for Health): Longitudinal Observational Study. Journal of Medical Internet Research. 2008;10(4):11.

204. Glasgow R, Lichtenstein E, Marcus A. Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness tradition. American Journal of Public Health. 2003;93(8):1261-67.

205. Kelders SM, Kok R, Ossebaard H, JE VG-P. Persausive system design does matter: a systematic review of adherence to web-based interventions. Journal of Medical Internet Research. 2012;14(6):e152.

206. Brueton V, Tierney J, Stenning S, Nazareth I, Meredith S, Hardin S, et al. Strategies to reduce attrition in randomised trials. Trials. 2011;12(Suppl 1):A128.

207. Khadjesari Z, Murray E, Kalaitzaki E, White I, McCambridge J, Thompson S, et al. Impact and costs of incentives to reduce attrition in online trials: two randomized controlled trials. Journal of Medical Internet Research. 2011;13(1):e26.

208. Buis LR, Poulton TA, Holleman RG, Sen A, Resnick PJ, Goodrich DE, et al. Evaluating Active U: an internet-mediated physical activity program. BMC Public Health. 2009;9:331.

209. Couper M, Alexander G, Zhang N, Little R, Maddy N, Nowak M, et al. Engagement and retention: measuring breadth and depth of participant use of an online intervention. Journal of Medical Internet Research. 2010;12(4):e52.

210. Anderson-Bill ES, Winett RA, Wojcik JR. Social cognitive determinants of nutrition and physical activity among web-health users enrolling in an online intervention: the influence of social support, self-efficacy, outcome expectations, and self-regulation. Journal of Medical Internet Research. 2011;13(1):e28.

211. Balmford J, Borland R, Benda P. Patterns of use of an automated interactive personalized coaching program for smoking cessation. Journal of Medical Internet Research. 2008;10(5):e54.

212. Graham AL, Bock BC, Cobb NK, Niaura R, Abrams DB. Characteristics of smokers reached and recruited to an Internet smoking cessation trial: a case of denominators. Nicotine and Tobacco Research. 2006;8(Suppl 1):S43-S8.

213. Graham AL, Milner P, Saul JE, Pfaff L. Online advertising as a public health and recruitment tool: comparison of different media campaigns to increase demand for smoking cessation interventions. Journal of Medical Internet Research. 2008;10(5):e50.

214. Wanner M, Martin-Diener E, Braun-Fahrlander C, Bauer G, Martin B. Effectiveness of Active-Online, an individually tailored physical activity intervention, in a real-life setting: randomized controlled trial. Journal of Medical Internet Research. 2009;11(3).

215. Hansen A, Bronbaek M, Helge J, Severin M, Curtis T, Tolstrup J. Effect of a web-based intervention to promote physical activity and improve health among physicially inactive adults: a population-based randomized controlled trial. Journal of Medical Internet Research. 2012;14(5):e145.

216. Vosbergen S, Laan E, Colkesen E, Niessen M, Kraaijenhagen R, Essink-Bot M, et al. Evaluation of end-user satisfaction among employees participating in a web-based health risk assessment with tailored feedback. Journal of Medical Internet Research. 2012;14(5):e140.

217. Roberts EB, Ramnath R, Fallows S, Sykes K. "First-hit" heart attack risk calculators on the world wide web: Implications for laypersons and healthcare practitioners. International Journal of Medical Informatics. 2008;77(6):405-12.

218. Whittaker R, Bramley D, Wells S, Stewart A, Selak V, Fuirness S, et al. Will a web-based cardiovascular disease (CVD) risk assessment programme increase the assessment of CVD risk factors for Maori? New Zealand Medical Journal. 2006;119(1238):U2077.

219. Buist D, Knight Ross N, Reid R, Grossman D. Electronic health risk assessment adoption in an integrated healthcare system. American Journal of Managed Care. 2014;20(1):62-9.

220. Kerr C, Murray E, Noble L, Morris R, Bottomley C, Stevenson F, et al. The potential of web-based interventions for heart disease self-management: a mixed methods investigation. Journal of Medical Internet Research. 2010;12(4):e56.

221. Harle C, Padman R, Downs J, editors. The impact of web-based diabetes risk calculators on information processing and risk perception. American Medical Informatics Association Annual Symposium Proceeding. 2008; 2008:283-287.

222. Holmberg C, Harttig U, Schulze MB, Boeing H. The potential of the Internet for health communication: the use of an interactive on-line tool for diabetes risk prediction. Patient Education and Counseling. 2011;83:106-12.

223. Brouwer W, Oenema A, Raat H, Crutzen R, de Nooijer J, de Vries Nanne K, et al. Characteristics of visitors and revisitors to an Internet-delivered computer-tailored lifestyle intervention implemented for use by the general public. Health Education Research. 2010;25(4):585-95.

224. Neufinger N, Cobain M, Newson R. Web-based self-assessment health tools: who are the users and what is the impact of missing input information? Journal of Medical Internet Research. 2014;16(9):e215.

225. Schoenbach VJ, Wagner EH, Berry WL. Health risk appraisal: review of evidence for effectiveness. HSR: Health Services Research. 1987;22(4):553-80.

226. Soler RE, Leeks KD, Razi S, Hopkins D, Griffith M, Aten A, et al. A systematic review of selected interventions for worksite health promotion, the assessment of health risks with feedback. American Journal of Preventive Medicine. 2010;38(2S):S237-S62.

227. Lopez-Gonzalez A, Aguilo A, Frontera M, Bennasar-Veny M, Campos I, Vicente-Herrero T, et al. Effectiveness of the Heart Age tool for improving modifiable cardiovascular risk factors in a Southern European population: a randomized trial. European Journal of Preventive Cardiology. 2014;Feb 3, 2014 (E pub ahead of print).

228. Bonner C, Jansen J, Newel B, Irwig L, Glasziou P, Doust J, et al. I don't believe it, but I'd better do something about it: patient experiences of online heart age risk calculators. Journal of Medical Internet Research. 2014;16(5):e120.

229. Harle CA, Downs JS, Padman R. A Clustering Approach to Segmenting Users of Internet-based Risk Calculators. Methods of Information in Medicine. 2011;50:244-52.

230. Zbib A, Hodgson C, Calderwood S. Can ehealth tools help organizations reach their target populations? Healthcare Management Forum. 2011;24(4):150-9.

231. Polgar S, Thomas SA. Introduction to Research in the Health Sciences. Fifth edition ed. Philadelphia: Churchill Livingston Elsevier; 2008.

232. Neufingerl N, Cobain M, Newson R. Web-based self-assessment health tools: who are the users and what is the impact of missing input information? Journal of Medical Internet Research. 2014;16(9):e215.

233. Navarro FH, Wilkins ST. A new perspective on consumer health Web use: "valuegraphic" profiles of health information seekers. Managed Care Quarterly. 2001;9(2):35-43.

234. Hand D, Manila H, Smyth P. Principles of Data Mining. Cambridge, MA: MIT Press; 2001.

235. Sutton C. Overview of Data Mining. Edinburgh: University of Edinburgh, School of Informatics; 2012 [Updated 2012; cited 2013 March 13]; Available from: http://www.inf.ed.ac.uk/teaching/courses/dme/2012/slides/datamining_intro4up.pdf.

236. Galt V. Crunch the numbers: Data analytics specialists mine market demand. Globe and Mail. 2013 Friday, March 29, 2013;Sect. Business.

237. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: the next frontier for innovation, competition, and productivity: McKinsey Global Institute, McKinsey & Company; 2011. Available from: www.mckinsey.com/mgl/publications/.

238. Zikopoulos P, Eaton C, deRoos D, Deutsch T, Lapis G. Understanding big data: analytics for enterprise class Hadoop and streaming data. New York: McGraw Hill; 2012. Available from: www.ibm.com/software/data/education/bookstore.

239. Adams M. Sex in the Snow, Canadian Social Values at the End of the Millennium. Toronto: Penguin; 1997.

240. Bakaric IR. Uncovering regional disparities - the use of factor and cluster analysis. Croatian Economic Survey. 2006;9:11-34.

241. Berge JM, Wall M, Bauer KW, Neumark-Sztainer D. Parenting characteristics in the home environment and adolescent overweight: a latent class analysis. Obesity. 2010;18:818-25.

242. Morris LA, Grossman R, Barkdoll G, Gordon E. A segmentational analysis of prescription drug information seeking. Medical Care. 1987;25(10):953-64.

243. Lloyd J, Doll H, Hawton K, Dutton WH, Geddes JR, Goodwin GM, et al. Internet gamblers: a latent class analysis of their behaviours and health experiences. Journal of Gambling Studies. 2010;26:387-99.

244. Gelbard R, Goldman O, Spiegler I. Investigating diversity of clustering methods: an empirical comparison. Data and Knowledge Engineering. 2007;63(1):155-66.

245. Eshghi A, Haughton D, Legrand P, Skaletsky M, Woolford S. Identifying groups: a comparison of methodologies. Journal of Data Science. 2011;9:271-91.

246. Haughton D, LeGrand P, Woolford S. Review of three latent class cluster analysis packages: Latent GOLD, poLCA, and MCLUST. American Statistician. 2009;63(1):81-91.

247. Danermark B, Ekstrom M, Jakobsen L, Karlsson JC. Explaining Society, Critical Realism in the Social Sciences. London and New York: Routledge; 2002.

248. Bhaskar R. Philosophy and scientific realism. In: Archer M, Bhaskar R, Collier A, Lawson T, Norrie A, editors. Critical Realism, Essential Readings. London and New York: Routledge; 1998.

249. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine. 1996;17(3):38-54.

250. Joffres MR, Hamer P, MacLean DR, Gilbert JL, Fodor JG. Distribution of blood pressure and hypertension in Canada and the United States. American Journal of Hypertension. 2001;14:1099-105.

251. Prochaska JO, Redding CA, Evers KE. The transtheoretical model and stages of change. In: Glanz K, Rimer BK, Lewis FM, editors. Health Behavior and Health Education, Theory, Research, and Practice. 3rd ed. ed. San Francisco: Jossey-Bass; 2002. p. 99-120.

252. Zimmerman GL, Olsen CG, Bosworth MF. A 'stages of change' approach to helping patients change behavior. American Family Physician. 2000;61(5):1409-16.

253. Richardson CG, Hamadani LG, Gotay C. Quantifying Canadians' use of the Internet as a source of information on behavioural risk factor modifications related to cancer prevention. Chronic Diseases and Injuries in Canada. 2013;33(3):123-8.

254. Nolan RP, Liu S, Shoemaker JK, Hachinski V, Lynn H, Mikulis DJ, et al. Therapeutic benefit of internet-based lifestyle counselling for hypertension. Canadian Journal of Cardiol. 2012;28:390-6.

255. Durrani S, Irvine J, Nolan RP. Psychosocial determinants of health behaviour change in an e-counseling intervention for hypertension. International Journal of Hypertension. 2012;Epub 2011 Dec 20.

256. Freenfield L. A definition of data warehousing. The Data Warehousing Information Center: LGI Systems Incorporated. [Updated no date; cited 2013 October 3]. Available from: http://www.dwinfocenter.org/defined.html.

257. Wentland E, Smith K. Survey Responses, An Evaluation of Their Validity. San Diego: Academic Press Inc.; 1993.

258. National Population Health Survey - Household Component – Cross-sectional (NPHS). Statistics Canada; 2007 [Updated 2007 October 24; cited 2014 November 11].

259. Rea L, Parker R. Designing and Conducting Survey Research, A Comprehensive Guide. 2nd edition ed. San Francisco: Jossey-Bass; 1997.

260. Barry M, Walker-Corkery E, Chang Y, Tyll L, Cherkin D, Fowler F. Measurement of overall and disease-specific health status: does the order of questionnaires make a difference? Journal of Health Services Research. 1996;1(1):20-7.

261. Siminski P. Order effect in batteries of questions. Quality and Quantity. 2008;42(4):477-90.

262. Salant P, Dillman D. How to Conduct Your Own Survey. Toronto: John Wiley and Sons; 1994.

263. Smith C, Fletcher J. Inside Information, Making Sense of Marketing Data. Chicester: Wiley; 2001.

264. Statistics Canada. Canadian Community Health Survey (CCHS). Annual Component - 2010 Questionnaire. Ottawa: Statistics Canada, 2011. [Updated 2012 June 12; cited 2014 November 3]. Available from: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SurvId=50653&InstaId=114112&SDDS=3226

265. Health Canada. Sodium in Canada. Health Canada; [updated 2012 July 12; cited 2014 June 28]. Available from: http://www.hc-sc.gc.ca/fn-an/nutrition/sodium/index-eng.php.

266. Butt P, Berirness D, Gliksman L, Paradis C, Stockwell T. Alcohol and Health in Canada: A Summary of Evidence and Guidelines for Low Risk Drinking. Ottawa: Canadian Centre on Substance Abuse; 2014 Available from: http://www.ccsa.ca/Eng/topics/alcohol/drinking-guidelines/Pages/default.aspx

267. Ma J, Betts NM, Horacek T, Georgiou C, White A. Assessing stages of change for fruit and vegetable intake in young adults: a combination of traditional staging alogirthms and food-frequency questionnaires. Health Education Research. 2003;18(2):224-36.

268. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?". Lancet. 2005;365:82-03.

269. Kelsey JL, Whittemore AS, Evans AS, Thompson WD. Methods in Observational Epidemiology (2nd ed). New York: Oxford University Press; 1996.

270. Newell S, Girgis A, Sanson-Fisher R, Savolained MJ. The accuracy of self-reported health behaviors and risk fators relating to cancer and cardiovascular disease in the general population: a ciritical review. American Journal of Preventive Medicine. 1999;17(3):211-29.

271. Wong S, Shields M, Leatherdale S, Malaison E, Hammond D. Assessment of validity of self-reported smoking status. Health Reports. 2012;23(1):1-7.

272. Klein JD, Thomas RK, Sutter EJ. Self-reported smoking in online surveys: prevalence estimate validity and item format effects. Medical Care. 2007;45(7):691-5.

273. Shields M, Connor Gorber S, Tremblay MS. Estimates of obesity based on self-report versus direct measures. Health Reports. 2008;19(2):61-76.

274. Elger FJ, Stewart JM. Validity of self-report sreening for overweight and obesity. Evidence from the Canadian Community Health Survey. Canadian Journal of Public Health. 2008;99(5):423-7.

275. Dal Grande E, Fullerton S, Taylor AW. Reliability of self-reported health risk factors and chronic conditions quesions collected using the telephone in South Australia, Australia. BMC Medical Research Methodology. 2012;12:108.

276. Martin L, Leff M, Calone N, Garrett C, Nelson D. Validation of a self-reported chronic condition and health services in a managed care population. American Journal of Preventive Medicine. 2000;18(3):215-18.

277. Gorber SC, Tremblay MS, Campbell N, Hardt J. The accuracy of self-reported hypertension: a systematic review and meta-analysis. Current Hypertension Reviews. 2014;10(4):35-62.

278. Janssens AC, Henneman L, Delmar SB, Khoury MJ, Steverberg EW, Eijkemans MJ, et al. Accuracy of self-reported family history is strongly influenced by the accuracy of self-reported personal health status of relatives. Journal of Clinical Epidemiology. 2012;65(1):82-9.

279. Van Eenwyk J, Bensley L, Ossiander EM, Krueger K. Comparison of examination-based and self-reported risk factors for cardiovascular disease, Washington State, 2007-2007. Preventing Chronic Disease. 2012;9(110321).

280. Leenen FHH, Dumais J, McInnis NH, Turton P, Stratychuk L, Nemeth K, et al. Results of the Ontario Survey on the Prevelance and Control of Hypertension. Canadian Medical Association Journal. 2008;178(11):1441-9.

281. Bensen JT, Liese AD, Rushing JT, Province M, Folsom AR, Rich SS, et al. Accuracy of proband reported family history: the NHLBI Family Heart Study (FHS). Genetic Epidemiology. 1999;17(2):141-50.

282. Grimm P. Social Desirability Bias. Wiley International Encyclopedia of Marketing: John Wiley & Sons, Ltd; 2010.

283. Hinz A, Michalski D, Schwarz R, Herzberg P. The acquiescence effect in responding to a questionnaire. GMS Psych-social Medicine. 2007;4:Doc07.

284. Greene J, Speizer H, Witala W. Telephone and web: mixed-mode challenge. Health Services Research. 2008;43(1 Pt 1):230-48.

285. Canadian Institute for Health Information. Reducing Gaps in Health: A Focus on Socio-Economic Status in Urban Canada. Ottawa: Canadian Institute for Health Information, 2008.

286. Mikkonen J, Raphael D. Social Determinants of Health: The Canadian Facts. Toronto: York University School of Health Policy and Management; 2010.

287. Wilkins R, Tjepkema M, Mustard C, Choiniere R. The Canadian census mortality follow-up study, 1991 through 2001. Health Reports. 2008;19(3):26-43.

288. Tjepkema M, Wilkins R, Long A. Cause-specific mortality by education in Canada: a 16-year follow-up study. Health Reports. 2012;23(3):3-11.

289. Mustard CA, Derksen S, Berthelot J-M, Wolfson M, Roos LL. Age-specific education and income gradients in morbidity and mortality in a Canadian province. Social Science & Medicine. 1997;45(3):383-97.

290. Steele LS, Dewa CS, Lin E, Lee KLK. Education level, income level and mental health services use in Canada: associations and policy implications. Healthcare Policy. 2007;3(1):06-106.

291. Setayeshgar S, Whiting SJ, Vatanparast H. Prevalence of 10-year risk of cardiovascular diseases and associated risks in Canadian adults: the contribution of cardiometabolic risk assessment introduction. International Journal of Hypertension. 2013;2013.

292. Manuel DG, Rosella LCA, Tuna M, Bennett C. How many Canadians will be diagnosed with diabetes between 2007 and 2017? Assessing the population risk. ICEES Investigative Report. Toronto: Institute for Clinical Evaluative Sciences, 2010.

293. Lindsay J, Laurin D, Verreault R, Hebert R, Helliwell B, Hill GB, et al. Risk factors for Alzheimer's disease: a prospective analysis from the Canadian Study of Health and Aging. American Journal of Epidemiology. 2002;156(5):445-53.

294. Mao Y, Hua J, Ugnat A-M, Semenciw R, Fincham S, and the Canadian Care Registries Epidemiology Research Group. Socioeconomic status and lung cancer risk in Canada. International Journal of Epidemiology. 2001;30(4):800-17.

295. Haydon E, Roerecke M, Giesbrecht N, Rehm J, Kobus-Matthews M. Chronic Disease in Ontario and Canada: Determinants, Risk Factors and Prevention Priorities. Toronto: Ontario Chronic Disease Prevention Alliance and the Ontario Public Health Association, 2006.

296. Choiniere R, Lafontaine P, Edwards AC. Distribution of cardiovascular disease risk factors by socioeconomic status among Canadian adults. Canadian Medical Association Journal. 2000;162(9 Suppl):S13-S24.

297. Kline RB. Beyond Significance Testing, Reforming Data Analysis Methods in Behavioral Research. Washington, DC: American Psychological Association; 2004.

298. Sackett DL, Haynes RB, Tugwell P. Clinical Epidemiology, A Basic Science for Clinical Medicine. Boston/Toronto: Little, Brown and Company; 1985.

299. Grissom RJ, Kim JJ. Effect Sizes for Research, A Broad Practical Approach. New York: Lawrence Erlbaum Associates; 2005.

300. Sheskin DJ. Handbook of Parametric and Nonparametric Statistical Procedures. Fourth edition. ed. Boca Faton, Florida: Chapman & Hall/CRC; 2007.

301. Field A. Discovering Statistics Using SPSS. Third Edition ed. London: Sage; 2009.

302. Pallant J. SPSS Survival Manual, A Step by Step Guide to Data Analysis using SPSS for Windows. Third edition ed. Maidenhead, England: Open University Press; 2007.

303. Crewson P. Applied Statistics Handbook, Version 1.2, Coefficients for Measuring Association. AcaStat Software; 2012 [Updated no date; cite 2012 July 13]. Available from: http://www.acastat.com/Statbook/chisqassoc.htm.

304. Norman GR, Treiner DL. Biostatistics, The Bare Essentials. St. Louis, Missouri: Mosby; 1994.

305. MacIagan L, Park J, Sanmartin C, Mathur K, Roth D, Manuel D, et al. The CANHEART health index: a tool for monitoring the cardiovascular health of the Canadian population. Canadian Medical Association Journal. 2014;186(3):180-7.

306. Statistics Canada. Table 109-5325. Estimates of population (2006 Census and administrative data) by age group and sex for July 1st, Canada, provinces, territories, health regions (2011 boundaries) and peer groups. Statistics Canada; 2012 [Updared 2014 June 12; cited 2012 March 4]; Available from: http://www5.statcan.gc.ca/cansim/a26.

307. Statistics Canada. Table 105-0502. Health indicator profile, two year period estimates, by age group and sex, Canada, provinces, territories, health regions (2011 boundaries) and peer groups.: Statistics Canada; 2011 [Update 2013 June 12; cited 2012 March 4]; Available from: http://www5.statcan.gc.ca/cansim/a26?lang=eng&retrLang=eng&id=1050502&tabMode=dataTable&srchLan=-1&p1=-1&p2=9.

308. Wanner M, Martin-Diener E, Bauer G, Braun-Fahrlander C, Martin B. Comparison of trial participants and open access users of a web-based physical activity intervention regarding adherence, attrition, and repeated participation. Journal of Medical Internet Research. 2010;12(1):e3.

309. Colkesen EB, Ferket BS, Tijssen JBP, Kraaijenhagen RA, van Kalken CK, Peters RJG. Effects on cardiovascular disease risk of a web-based health risk assessment with tailored health advice: a follow-up study. Vascular Health and Risk Management. 2011;7:67-74.

310. Schulz DN, Smit ES, Stanczyk NE, Kremer S, de Vries H, Evers SMAA. Economic evaluation of a web-based tailored lifestyle intervention for adults: findings regarding cost-effectiveness and cost-utility from a randomized controlled trial. Journal of Medical Internet Research. 2014;16(3):e91.

311. Viera AJ. Odds ratios and risk ratios: what's the difference and why does it matter? Southern Medical Association. 2008;101(7):730-4.

312. Olivier J, Bell ML. Effect Sizes for 2×2 Contingency Tables. PLoS One. 2013;8(3):e58777.

313. Ferguson CJ. An effect size primer: a guide for clinicians and researchers. Professional Psychology: Research and Practice. 2009;40(5):532-8.

314. Hernan MA, Hernandez-Diaz S, Werier MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. American Journal of Epidemiology. 2002;155(2):176-84.

315. Baily KD. Typologies and Taxonomies, An Introduction to Classification Techniques. Thousand Oaks, California: Sage Publications; 1994.

316. Lustria MLA, Cortese J, Noar SM, Glueckaluf RL. Computer-tailored health interventions delivered over the web: Review and analysis of key components. Patient Education and Counseling. 2009;74(2):156-73.

317. Hawkins R, Kreuter MW, Resnicow K, Fishbein M, Dijkstra A. Understanding tailoring in communicating about health. Health Educ Res. 2008;23(3):454-66.

318. Lewis M, McCormack L. The intersection between tailored health communication and branding for health promotion. In: Evans W, Hasting G, editors. Public Health Branding: Applying Marketing for Social Change. Oxford: Oxford University Press; 2008.

319. Lefebvre RC, Flora J. Social marketing and public health interventions. Health Education Quarterly. 1988;15(3):299-315.

320. Slater M. Theory and method in health audience segmentation. Journal of Health Communication: International Perspectives. 1996;1(3):267-83.

321. Albrecht T. Advances in segmentation modeling for health communication and social marketing campaigns. Journal of Health Communication: International Perspectives. 1996;1(1):65-80.

322. Noar S, Benae C, Harris M. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychological Bulletin. 2007;133(4):673-93.

323. Leventhal A, Huh J, Dunton G. Clustering of modifiable biobehavioral risk factors for chronic disease in US adults: a latent class analysis. Perspectives in Public Health. 2014;134(6):331-8.

324. Williams J, Flora J. Health behavior segmentation and campaign planning to reduce cardiovascular disease risk among Hispanics. Health Education Quarterly. 1995;22(1):36-46.

325. Maibach E, Weber D, Massett HA, Hancock G, Price S. Understanding consumers' health information preferences: development and validation of a brief screening instrument. Journal of Health Communication: International Perspectives. 2006;11(8):717-36.

326. The U.S. Health Care Market: A Strategic View of Consumer Segmentation. Washington, DC: Deloitte Center for Health Solutions, 2012. Available from: http://www2.deloitte.com/us/en/pages/life-sciences-and-health-care/articles/center-for-health-solutions-survey-of-us-consumers-health-care.html

327. Skinner C, Campbell MK, Rimer BK, Curry S, Prochaska J. How effective is tailored print communication? Annals of Behavioral Medicine. 1999;21(4):290-98.

328. Rimer BK, Kreuter MW. Advancing tailored health communication: a persuasion and message effects perspective. Journal of Communication. 2006;56(Supplement 1):S184-S201.

329. Dijkstra A. The psychology of tailoring-ingredients in computer-tailored persuasion. Social and Personality Psychology Compass. 2008;2(2):765-84.

330. Center for Disease Control and Prevention. What Markes Social Marketing Different? Audience Segmentation. Social Marketing Basics: Center for Disease Control and Prevention; [Updated no date; cited 2014 November 11]. Available from: http://www.cdc.gov/nccdphp/dnpa/socialmarketing/training/basics/audience_segmentation.htm.

331. Romesburg HC. Cluster Analysis for Researchers. North Carolina: Lulu Press; 2004.

332. Aldenderfer MS, Blashfield RK. Cluster Analysis. Thousand Oaks, California: Sage Publications; 1984.

333. Dolnicar S. Empirical market segmentation: what you see is what you get. In: Theobald W, editor. Global Tourism, the Next Decade. 3rd ed. Oxford: Butterworth-Heinemann; 2005. p. 309-25.

334. Field A. Cluster Analysis (handout). www.statisticalshell.com. 2000 [Update 2000 February 3; cited 2012 May 29]. Available from: http://www.statisticshell.com/docs/cluster.pdf.

335. Tan P-N, Steinbach M, Kumar V. Introduction to Data Mining. Boston: Addison-Wesley; 2006. p.

336. Norusis MJ. IBM SPSS Statistics 19 Statistical Procedure Companion. Boston: Addison Wesley; 2011.

337. The SPSS TwoStep Cluster Component, A scalable component enabling more efficient customer segmentation. White paper- technical report. SPSS Corporation; Armonk, New York; 2001. Available from: http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf

338. Uebersax J. LCA Frequently Asked Questions (FAQ). John Uebersax Enterprises; 2009 [Updated 2009 July 8; cited 2012 October 10]. Available from: http://www.john-uebersax.com/stat/faq.htm.

339. McCutcheon AL. Latent Class Analysis. Thousand Oaks, California: Sage Publications; 1987.

340. Mooi E, Sarstedt M. A Concise Guide to Market Research, The Process, Data, and Methods Using IBM SPSS Statistics. New York: Springer; 2011.

341. Dolnicar S. Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement. Australiasian Journal of Market Research. 2003;11(2):5-12.

342. Sambandam R. Cluster analysis gets complicated Marketing Research. 2003;15(1):16-21.

343. Stockburger DW. Cluster Analysis. Multivariate Statistics: Concepts, Models and Applications: Missouri State University; 1998.

344. Anderson L, Weiner JL. Actionable Market Segmentation Guarenteed, A White Paper from the Ipsos Group. Market Segmentation2004.

345. Steptoe A, Kerry S, Rink E, Hilton S. The impact of behavioral counseling on stage of change in fat intake, physical activity, and cigarette smoking in adults at increased risk of coronary heart disease. American Journal of Public Health. 2001;91(2):265-9.

346. Ronda B, Van Assema P, Brug J. Stages of change, psychological factors and awareness of physical activity levels in the Netherlands. Health Promotion International. 2001;16(4):305-413.

347. Ewing Garber C, Allsworth J, Marcus B, Hesser J, Lapine K. Correlates of the stages of change for physical activity in a population survey. American Journal of Public Health. 2008;98(5):897-904.

348. Higgins O, Sixsmith J, Barry M, Domegan C. A literature review of health information-seeking behaviour on the web: a health consumer and health professional perspective. Stockholm: EDIC (European Centre for Disease Prevention and Control), 2011.

349. Public Health Agency of Canada, Canadian Institute for Health Information, Canadian Stroke Network, Heart and Stroke Foundation of Canada, Statistics Canada. Tracking Heart Disease and Stroke in Canada. Ottawa: 2009 Report: HP3203/2009E.

350. Statistics Canada. Heart health and choletserol levels of Canadians, 2007 to 2009. Ottawa: Statistics Canada, March 2010.

351. Wilkins K, Campbell NRC, Joffres MR, McAlister FA, Nichol M, Quach S, et al. Blood pressure in Canadian adults. Ottawa: Statistics Canada, 2010 Report No: 82-003-X.

352. Maletta H. Paper on Weighting. SPSS Tutorial, Raynald's SPSS Tools [Updated 2012 March 31; Cited 2014 November 12]. Available from: http://www.spsstools.net/spss.htm

353. Arthritis Community Research and Evaluation Unit (ACREU) for the Arthritis Society. Arthritis in Canada. Toronto: Arthritis Society, 2013.

354. Heart and Stroke Foundation of Canada. Tipping the Scales of Progress, Heart Disease and Stroke in Canada 2006. Ottawa: Heart and Stroke Foundation of Canada, 2006.

355. Canadian Diabetes Association and Diabetes Quebec. Diabetes: Canada at the Tipping Point, Charting a New Path. Toronto: Canadian Diabetes Association, 2011.

356. Centers for Disease Control and Prevention. Table 2-1 Lifetime asthma prevalence percents by age, United States: National Health Interview Survey, 2009. Atlanta: Centers for Diseas Control and Prevention; 2010 [updated 2010 October 1; cited 2013 August 2]. Available from: http://www.cdc.gov/asthma/nhis/09/table2-1.htm.

357. MacLeod J, Smith G, Heslop P, Metcalfe C, Carroll D, Hart C. Psychological stress and cardiovascular disease: empirical demonstration of bias in a prospective observational study of Scottish men. BMJ. 2002;324:1247.

358. Greenwood C, Muir K, Packham C, Madeley R. Coronary heart disease: a review of the role of psychosocial stress and social support. Journal of Public Health Medicine. 1996;18(2):221-31.

359. Ariyo A, Haan M, Tangen C, Rutledge J, Cushman M, Dobs A, et al. Depressive symptoms and risks of coronary heart disease and mortality in elderly Americans. Circulation. 2000;102:1773-9.

360. Fiedorowicz J, He J, Merikangas K. The association between mood and anxiety disorders with vascular diseases and risk factors in a nationally representative sample. Journal of Psychosomatic Research. 2011;70(2):145-54.

361. Patten S, Wang J, Williams J, Currie S, Beck C, Maxwell C, et al. Descriptive epidemiology of major depression in Canada. Canadian Journal of Psychiatry. 2006;51(2):84-90.

362. Khadjesari Z, Murray E, Kalaitzaki E, White I, McCambridge J, Thompson S, et al. Impact and costs of incentives to reduce attrition in online trials: two randomized controlled trials. Journal of Medical Internet Research. 2011;13(1):e26.

363. Boughner R. Volunteer Bias. In: Salkind M, editor. Encyclopedia of Research Design. Thousand Oaks: SAGE Publications; 2010. p. 1609-11.

364. Almeida L, Kashden T, Nunes T, Coelho R, Albin-Teixeira A, Soares-da-Silva P. Who volunteers for phase I clinical trials? Influences of anxiety, social anxiety and depressive symptoms on self-selection and the reporting of adverse events. European Journal of Clinical Pharmacology. 2008;64(6):575-82.

365. Golomb B, Chan V, Evans M, Koperski S, White H, Criqui M. The older the better: are elderly study participants more non-representative? A cross-sectional analysis of clinical trial and observational study samples. BMJ Open. 2012;2:e000833.

366. Eysenbach G, Wyatt J. Using the internet for surveys and health research. Journal of Medical Internet Research. 2002;4(2):e13.

367. Thompson C. If you could just provide me with a sample: examining sampling in qualitative and quantitative research papers. Evidence-based Nursing. 1999;2(3):68-70.

368. Peels D, Bolman C, Golstein R, De Vries H, Mudde A, Vvan Stralan M, et al. Differences in reach and attrition between web-based and print-delivered

tailored intervention among adults over 50 years of age: clustered randomized trial. Journal of Medical Internet Research. 2012;14(6):e179.

369. Robroek S, Brouwer W, Lindeboom D, Oenema A, Burdorf A. Demographic, behavioral, and psychosocial correlates of using the website component of a worksite physical activity and healthy nutrition promotion program: a longitudinal study. Journal of Medical Internet Research. 2010;12(3):e4.

370. Jousilahti P, Puska P, Vartiainen E, Pekkanen J, Tuomilehto J. Parental history of premature coronary heart disease: An independent risk factor of myocardial infarction. Journal of clinical epidemiology. 1996;49(5):497-503.

371. Magidon J, Vermunt JK. Latent Class Models. In: Kaplan D, editor. The Sage Handbook of Quantitative Methodology for the Social Sciences. Thousand Oaks: Sage Publications; 2004. p. 175-98.

372. Newsom JT, Huguet N, Ramage-Morin PL, McCarthy MJ, Bernier J, Kapolan MS, et al. Health behaviour changes after diagnosis of chronic illness among Canadians aged 50 or older. Health Reports. 2012;23(4):1-7.

373. Jackson J. Data mining: a conceptual overview. Communications of the Association for Information Systems. 2002;8:267-96.

374. Williams GC. Improving patients' health through supporting the autonomy of patients and providers. In: Deci EL, Ryan RM, editors. Handbook of Self-Determination Research. Rochester, NY: University of Rochester Press; 2002. p. 233-54.

375. Sheldon KM, Williams GC, Joiner T. Self-Determination Theory in the Clinic, Motivating Physical and Mental Health. New Haven, CT: Yale University Press; 2003.

376. National Institute for Health and Clinical Excellence. Behaviour Change at Population, Community and Individual Levels. London: NICE, 2007.

377. Krotoski A. Tech Weekly: Email overload London: The Guardian; 2010 [cited 2012 August 24]. Available from: http://www.guardian.co.uk/technology/blog/audio/2010/sep/21/email-xobni-haystack-diaspora.

378. Whittaker S, Sidner C. Email overload: exploring personal information management of email. CHI ’96 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1996; 276-283.

379. Haynes GA. Testing the boundaries of the choice overload phenomenon: the effect of number of options and time pressure on decision difficulty and satisfaction. Psychology & Marketing. 2009;26:204-12.

380. Jessup RK, Veinott ES, Todd PM, Busemeyer JR. Leaving the store empty-handed: testing explanations for the too-much-choice effect using decision field theory. Psychology & Marketing. 2009;26:299-320.

381. Haubl G, Trifts V. Consumer decision making in online shopping environments: the effects of interactive decision aids. Marketing Science. 2000;19(1):4-21.

382. Statistics Canada. Generations in Canada, Age and sex, 2011 Census. Ottawa: Statistics Canada, 2012. Catalogue no. 98-311-X2011003.

383. Marketing Charts staff. Baby Boomers Control 70% of US Disposable Income. Thetford Center, Vermont: Marketing Charts; 2012 [Updated 2012 August 7; cited 2013 March 3]; Available from: http://www.marketingcharts.com/television/baby-boomers-control-70-of-us-disposable-income-22891/.

384. White D. Across the Divide, Tacking Digitial Exclusion in Glasgow. Dunfemline: Carnegie UK Trust, 2013.

385. Adler NE, Ostrove JM. Socioeconomic status and health: what we know and what we don't know In: Adler NE, Marmot M, McEwen BS, Stewart J, editors. Socioeconomic Status and Health in Industrial Nations, Social, Psychological, and Biological Pathways. New York, New York: New York Academy of Science; 1999. p. 3-15.

386. Evans RG, Stoddart GL. Producing health, consuming health care. In: Evans RG, Barer ML, Marmot TR, editors. Why Are Some People Healthy and Others Not? The Determinants of Health of Populations. Hawthorne, NY: Aldine de Gruyter; 1994. p. 27-64.

387. Marton C, Choo CW. A review of theoretical models of health information seeking on the web. Journal of Documentation. 2012;68(2):330-52.

388. Gallo L, Smith T, Cox C. Socioeconomic status, psychosocial processes, and perceived health: an interpersonal perspective. Annals of Behavioral Medicine. 2006;31(2):198-9.

389. Aujoulat I. Reconsidering patient empowerment in chronic illness: a critique of models of self-efficacy and bodily control. Social Science & Medicine. 2008;66:1228-39.

390. Audulv A, Asplund K, Norbergh KG. Who's in charge? The role of responsibility attribution in self-management among people with chronic illness. Patient Education and Counseling. 2010;81(1):94-100.

391. Archer MS. Making our Way Through the World, Human Reflexivity and Social Mobility. Cambridge: Cambridge University Press; 2007.

392. Evans W. How social marketing works in health care. BMJ. 2006;332(7551):1207-10.

393. Gallant M, Dorn G. Gender and race differences in the predictors of daily health practices among older adults. Health Education Research. 2001;16(1):21-31.

394. Kreuter MW, Skinner C. Tailoring: what's in a name? Health Education Research. 2000;15(1):1-4.

395. New Parks Associates Digital Health Research Identifies Four Consumer Health Groups. Dallas: Connected Health Summit; 2014 [Update 2014 August 7; cited 2014 November 18]. Available from: http://www.parksassociates.com/events/connected-health/media/chs-2014-pr6.

396. Chirrey S, Hodgson C. My CancerIQ A new online cancer prevention tool from Cancer Care Ontario: Opportunities and applications for public health. Toronto: Public Health Ontario; 2014 [Updated 2014 November 19; cited 2014 December 3]. Available from: http://www.publichealthontario.ca/en/LearningAndDevelopment/Events/Documents/My_CancerIQ_online_cancer_prevention_tool_Chirrey_Hodgson_2014.pdf.

397. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Computing Surveys. 1999;31(3):264-323.

398. Abbas OA. Comparisons between data clustering algorithms. International Arab Journal of Information Technology. 2008;5(3):320-5.

399. Hammouda K, Karray F, editors. A comparative study of data clustering techniques. Ninth SIAM International Conference on Data Mining; 2009 April

30 - May 2, 2009; Sparks, Nevada: Society for Industrial and Applied Mathematics.

400. Bartholomew D, Knott M, Moustaki I. Latent Variable Models and Factor Analysis, A Unified Approach. 3rd ed. ed. West Sussex: Wiley; 2011.

401. Uebersax J. LCA Software. 2012 [Updated 2012 May 10; cited 2012 August 24]; Available from: http://www.john-uebersax.com/stat/soft.htm.

402. Liu C, Liu Y-H, Xo T. To search is to believe? A comparative study of health information use by internet users. Proceedings of the American Society for Information Science and Technology. 2009;46(1):1-5.

403. Jacobs N, De Bourdeaudhuij I, Claes N. Surfing depth on a behaviour change website: predictors and effects on behaviour. Informatics for Health and Social Care. 2010;35(2):41-52.

404. Chapman LS, Rowe D, Witte K. eHealth Portals: who uses them and why? American Journal of Health Promotion. 2010;24(5):TAHP-1-7.

405. Riedesel P. Applying archetypal analysis in marketing research. Minneapolis: Action Marketing Research; 2008 [Updated 2008; cited 2012 December 24]. Available from: http://www.action-research.com/aaa.pdf.

406. Schwarzer R. Modeling health behavior change: How to predict and modify the adoption and maintenance of health behaviors. Applied Psychology: An International Review. 2008;57:1-29.

407. Sniehotta F, Scholz U, Schwarzer R. Bridging the intetion-behaviour gap: Planning, self-efficacy, and action control in the adoption and maintenance of physical exercise. Psychology & Health. 2005;20(2):143-60.

408. Sutton S. How does the Health Action Process Approach (HAPA) bridge the intention-behavior gap? An examination of the model's causal structure. Applied Psychology: An International Review. 2008;57(1):66-74.

409. Griffiths J, Blair-Steven C, Thorpe A. Social Marketing for Health and Specialized Health Promotion. Stronger Together - Weaker Apart. A paper for debate. London: Royal Society for Public Health and National Social Marketing Centre, 2008.

410. Segmenting Audiences to Promote Energy Balance, Resource Guide for Public Health Professionals. Betheseda, Maryland: Department of Health and Human Services, Centers for Disease Control and Prevention, n.d.

411. Keller H. Top 3 reasons to segment your audience.TEDx Montlake Cut; 2013 [Updated 2013; cited 2014 April 30]. Available from: http://www.youtube.com/watch?v=hsVRlzRNErY

412. Work Group for Community Health and Development. Community Tool Box, Section 4: Segmenting the Marget to Reach the Target Population. Lawrence, KS: University of Kansas; 2013 [Updated 2014; cited 2014 June 28]. Available from: http://ctb.ku.edu/en/table-of-contents/sustain/social-marketing/reach-targeted-populations/main

413. Crompton S. Off-reserve Aboriginal internet users. Canadian Social Trends. 2004;Winter:8-14.

414. Lefebvre RC. Segmentation: The first critical decision. South Florida University: On Social Marketing and Social Change (blog); 2005 [Updated 2008 August 20; cited 2014 April 28]. Available from: http://socialmarketing.blogs.com/r_craiig_lefebvres_social/2005/12/segmentation_th.html.

415. Reading C, Wien F. Health Inequalities and Social Determinants of Aboriginal Peoples' Health. Prince George, B.C.: National Collaborating Centre for Aboriginal Health, 2009.

416. Waterloo Wellington LHIN. Chronic conditions in the Waterloo Wellington LHIN. Waterloo, Ont: Waterloo Wellington Local Health Integration Network 2007.

417. Penna RM. The NonProfit Outcomes Toolbox, A Complete Guie to Program Effectiveness, Performance Measurement, and Results. Hoboken, NJ: John Wiley & Sons; 2011.

418. McMillan SJ. Health communication and the internet: relations between interactive characteristics of the medium and site creators, content, and purpose. Health Communication. 1999;11(4):375-90.

419. Rose G. Sick individuals and sick populations. International Journal of Epidemiology. 1985;14:32-8.

420. Cooney M-T, Dudina A, Whincup P, Capewell S, Menotti A, Jousilahti P, et al. Re-evaluating the Rose approach: comparative benefits of the population and high-risk prevention strategies. European Journal of Cardiovascular Prevention and Rehabilitation. 2009;16:541-9.

421. Redbirdonline. Health Promotion and Digital Channels: A New Framework for Successful Health Promotion Campaigns Victoria, BC: Redbird Communications; 2013 [Updated no date; cited 2014 November 21]. Available from: http://www.redbirdonline.com/sites/default/files/imce/Redbird%27s%20Guide%20to%20Digital%20Health%20Promotion%20.pdf.

422. Booth-Kewley S, Vickers RR. Associations between major domains of personality and health behavior. Journal of Personality 1994;62(3):281-98.

423. Krause KJ. Self-reported health: potential life underwriting tool? Journal of Insurance Medicine. 2002;34:61-7.

Appendix 1: Previous publication

Zbib A, Hodgson C. Calderwood S. Can eHealth tools enable health organizations to reach their target audience? *Healthcare Management Forum* 2011; 24: 155-159. Available at http://hmf.sagepub.com/content/24/3/155.full.pdf; Last accessed 8/05/2015.

Appendix 2: Heart&Stroke Risk Assessment Questionnaire

<table>
<tr><th>Function/ Question</th><th>Question</th></tr>
<tr><td>Consent</td><td>The Heart and Stroke Foundation is concerned about your privacy. All data you enter in this health assessment is considered confidential and will be treated according to the Heart and Stroke Foundation's Privacy Policy. Click here [LINK] for more information.</td></tr>
<tr><td>Preamble to Non-Modifiable Factors</td><td>Let's start by looking at factors that affect your health that you cannot control or change, such as your family history, whether you are a male or female, your age and your ethnic background.</td></tr>
<tr><td>Family History</td><td>Some health problems are more common in some families. Please tell us if you have a history of the following among your blood relatives. A blood relative refers to your natural or biological parents, grandparents, brothers, sisters, or children.
<table>
<tr><th>I have a blood relative with:</th><th>Yes</th><th>No</th><th>Don't Know</th></tr>
<tr><td>Diabetes or high blood sugar</td><td>( )</td><td>( )</td><td>( )</td></tr>
<tr><td>Heart disease:
• a female relative (grandmother, mother, sister or daughter) before she was age 65
• a male relative (grandfather, father, brother or son) before he was age 55</td><td>( )
( )</td><td>( )
( )</td><td>( )
( )</td></tr>
<tr><td>Stroke prior to age 65</td><td>( )</td><td>( )</td><td>( )</td></tr>
<tr><td>High blood pressure (hypertension)</td><td>( )</td><td>( )</td><td>( )</td></tr>
<tr><td>High cholesterol (hypercholesterolemia), an unhealthy cholesterol profile (dyslipidemia) or high triglycerides (another form of fat cell in the blood)</td><td>( )</td><td>( )</td><td>( )</td></tr>
</table></td></tr>
<tr><td>Gender</td><td>Some diseases are more common in men or women. Are you male or female?
( ) Male
( ) Female</td></tr>
<tr><td>Age
YOB [xxxx]; used to calculate AGE [current year-</td><td>Age can affect your risk of many diseases. What is your year of birth?
____

[If age < 20, then AGEDISCLAIMER and do not give HSF Recommendation (not eligible for BP On Track or HWAP; else, calculate risk status:</td></tr>
<tr><td>Ethnicity</td><td>Some ethnic groups are at higher or lower risk than others for some health problems. Please check the one ethnic group that you most see yourself as belonging to.
( ) Chinese [1]
( ) South Asian (e. g. Indian, Pakistani, Sri Lankan, Bangladeshi etc.) [2]
( ) Aboriginal North American (First Nations, Inuit, Metis) [3]</td></tr>
</table>

Function/ Question	Question
	( ) African heritage [4] ( ) Filipino [5] ( ) Southeast Asian (e.g. Vietnamese, Cambodian, Malaysian, etc.) [6] ( ) Japanese [7] ( ) Korean [8] ( ) Arab [9] ( ) West Asian (e.g. Iranian, Afghan etc.) [10] ( ) Latin American [11] ( ) White Caucasian [12] ( ) Other, please specify: [TEXT BOX] [13]
Preamble to Modifiable Factors	The following questions will now look at risk factors or conditions that you can change, treat or help to control.
Physical Activity	Are you moderately active at work or at home for at least 30 to 60 minutes, 4 or more days of the week? “Moderate” activity means such things as brisk walking, active gardening, swimming, dancing or biking. ( ) Yes, for more than six months [1] ( ) Yes but for less than six months [2] ( ) No but I’d like to start becoming more active within the next 30 days [3] ( ) No but I’d like to start becoming more active within the next 6 months [4] ( ) No and I do not plan on becoming more active [5]
Smoking	Do you smoke? ( ) No [1] ( ) Yes but I’ve already started trying to quit or cut down [2] ( ) Yes but I’d like to stop smoking within the next 30 days [3] ( ) Yes but I’d like to stop smoking within the next 6 months [4] ( ) Yes and I do not plan to stop smoking [5]
Body Mass Index (BMI)	The Body Mass Index (BMI) is one measure of your weight. To learn your BMI, please fill in your height and weight. Weight: _____ lbs OR _____ kg Height: _____ feet ___ inches OR _____ cm ( ) I don’t know my height or weight [BMI categories <18.5 (underweight) = 1 18.5-24.9 (normal weight) = 2 25.0 – 29.9 (overweight) = 3 ≥30.0 (obese) = 4]
Waist circumference (WC)	Your risk of some health problems, such as high blood pressure, heart disease, stroke, diabetes and high cholesterol, is affected how much weight you carry around your middle. Please enter your waist measurement. If you need help on how to measure your waist, please click here. [LINK] My waist measurement is: ___ inch or ___ cm OR ( ) I don’t know my waist measurement If gender = male, ETH = South Asian or Chinese and WC <90 cm (3 in) then low risk [1] If gender = male and ETH <> South Asian or Chinese and WC <102 (40 in) then low risk [1] If gender = male, ETH = South Asian or Chinese and WC ≥90 cm (3

Function/ Question	Question
	in) then high risk [2] If gender = male and ETH <> South Asian or Chinese and WC ≥102 (40 in) then high risk [2] If gender = female, ETH = South Asian or Chinese and WC <80 cm (32 in) then low risk [1] If gender = female, ETH <> South Asian or Chinese and WC < 88 cm (35 in) then low risk [1] If gender = female, ETH = South Asian or Chinese and WC ≥80 cm (32 in) then high risk [2] If gender = female, ETH <> South Asian or Chinese and WC ≥ 88 cm (35 in) then high risk [2]
Weight readiness to change – only displayed if applicable (WGTR) [1, 2, 3 or 4]	(Show only if age > 18 and BMI >24.9 and/or waist indicates high risk; if falls into this category score +10 for HWAP) Your BMI or waist measurement suggests you could benefit from losing some weight. When would you be willing to make changes to get to a healthier weight? ( ) I'm already working on trying to get to a healthier weight [1] ( ) Within the next 30 days [2] ( ) Within the next 6 months [3] ( ) I'm not planning to make changes or lose weight [4]
Salt	Which of the following statements BEST describes the amount of salt in your diet? ( ) I eat a lot of prepared or canned foods that are high in salt and I like to salt my food at the table. [1] ( ) I don't pay any attention to the amount of salt in the foods I eat [2] ( ) I make a conscious effort to limit the amount of salt in my diet, such as not salting my food at the table and choosing reduced-sodium foods whenever possible. [3]
Salt readiness to change -- for those who indicate a high salt diet	[Only display if SALT = eat a lot or don't pay attention] When would you be willing to make changes to reduce the amount of salt in your diet? ( ) I'm already trying to monitor and reduce the amount of salt in my diet [1] ( ) Within the next 30 days [2] ( ) Within the next 6 months [3] ( ) I'm not willing to change the amount of salt in my diet. [4]
Alcohol	Typically, do you drink more than 1 or 2 drinks that contain alcohol a day, to a weekly maximum of 14 drinks for men or 9 drinks for women? One drink is equal to: 12 oz/341 mL of beer (5% alcohol), as in one bottle of beer 5 oz/142 mL of wine (12% alcohol), as in one glass of wine 1.5 oz/43 mL of spirits or hard liquor (40% alcohol), as in one shot of hard liquor ( ) No [1] ( ) No, because within the past 6 months I've reduce the amount I drink or have stopped drinking alcohol[2] ( ) Yes I do, but I'm willing to start decreasing the amount I drink within the next 30 days [3] ( ) Yes I do, but in the next 6 months I'd like to stop drinking this much [4] ( ) Yes I do and I don't plan to change my drinking habits. [5]

<table>
<tr><th>Function/ Question</th><th>Question</th></tr>
<tr><td>Nutrition/Diet question (NU)

FATTYF [1, 2 or 3]
FASTF [1, 2 or 3]
FISH [1, 2 or 3]
VEGFR [1, 2 or 3]</td><td>In a typical week, how frequently do you do the following?

<table>
<tr><th></th><th>Less than once a week [1]</th><th>1-2 times a week [2]</th><th>3 or more times a week [3]</th></tr>
<tr><td>Eat high fat foods such as whole dairy products, fatty meats, donuts, cookies, or deep-fried foods such as battered fish, fish and chips, samosas or Jamaican patties?</td><td>( )</td><td>( )</td><td>( )</td></tr>
<tr><td>Eat fast food items such as hamburgers, French fries or onion rings?</td><td>( )</td><td>( )</td><td>( )</td></tr>
<tr><td>Eat broiled, baked or poached fish (any kind of fish that is not deep-fried or battered)?</td><td>( )</td><td>( )</td><td>( )</td></tr>
<tr><td>Eat five or more servings of vegetables and fruits each day? One serving is equivalent to: one medium apple, banana or orange, 1 cup of raw leafy vegetables such as spinach or lettuce, ½ cup of cooked vegetables, ½ cup of chopped, cooked or canned fruit, or ¾ cup vegetable or fruit juice.</td><td>( )</td><td>( )</td><td>( )</td></tr>
</table></td></tr>
<tr><td>Readiness to change diet question
(</td><td>[show only if Diet qualifies as a risk factor]
Your answers suggest that you could benefit from eating a healthier diet. When would you be willing to make changes to eat a healthier diet?
( ) I’m already working on eating a healthier diet [1]
( ) Within the next 30 days [2]
( ) Within the next 6 months [3]
( ) I don’t plan to change my eating habits [4]</td></tr>
<tr><td>Stress</td><td>In a typical week, how frequently do you feel overwhelmed or stressed by the demands on you?
( ) Seldom or never [1]
( ) A few times [2]
( ) Often or most of the time [3]</td></tr>
<tr><td>Stress readiness</td><td>Are you interested in making changes to help you manage your stress better?
( ) Yes, and I’m already trying to make some changes [1]
( ) Yes, and I’d like to start making changes within the next 30 days [2]
( ) Yes, and I’d like to start making changes within the next 6 months [3]
( ) No, I don’t plan to manage my stress differently [4]</td></tr>
<tr><td>Chronic Conditions</td><td>Has a doctor or other healthcare professional ever told you that you have any of the following chronic (long-term) conditions? For each health condition you report, please tell us whether you are taking a prescription medication for it (a drug or treatment prescribed by a doctor or nurse).

<table>
<tr><td></td><td>I have been told by a health provider that I</td><td>My healthcare</td></tr>
</table></td></tr>
</table>

<table>
<tr><th>Function/ Question</th><th>Question</th></tr>
<tr><td></td><td>
<table>
<tr><td>(Yes)</td><td>have:</td><td>provider prescribed medication for this condition (check all that apply).</td></tr>
<tr><td>( )</td><td>Alzheimer’s disease or other form of dementia</td><td>( )</td></tr>
<tr><td>( )</td><td>Arthritis</td><td>( )</td></tr>
<tr><td>( )</td><td>Asthma</td><td>( )</td></tr>
<tr><td>( )</td><td>Cancer (any form)</td><td>( )</td></tr>
<tr><td>( )</td><td>Chronic Obstructive Pulmonary Disease (COPD) such as chronic bronchitis or emphysema</td><td>( )</td></tr>
<tr><td>( )</td><td>Chronic back pain</td><td>( )</td></tr>
<tr><td>( )</td><td>Depression or anxiety</td><td>( )</td></tr>
<tr><td>( )</td><td>Diabetes (type 1 or 2)</td><td>( )</td></tr>
<tr><td>( )</td><td>Heart attack or heart disease</td><td>( )</td></tr>
<tr><td>( )</td><td>High blood pressure (hypertension)</td><td>( )</td></tr>
<tr><td>( )</td><td>High cholesterol (hypercholesteremia), an unhealthy cholesterol profile (dyslipidemia) or high triglycerides (another form of fat cell in the blood)</td><td>( )</td></tr>
<tr><td>( )</td><td>Kidney (renal) disease</td><td>( )</td></tr>
<tr><td>( )</td><td>Liver disease</td><td>( )</td></tr>
<tr><td>( )</td><td>Osteoporosis (bone-thinning)</td><td>( )</td></tr>
<tr><td>( )</td><td>Sleep apnea (while you sleep, you frequently stop breathing for short periods of time)</td><td>( )</td></tr>
<tr><td>( )</td><td>Stroke or “mini-stroke” (transient ischemic attack or TIA)</td><td>( )</td></tr>
<tr><td>( )</td><td>Other chronic health condition, please specify: [TEXT BOX]</td><td>( )</td></tr>
</table>
</td></tr>
<tr><td>Medication compliance</td><td>[If report a medication for any of the conditions listed, ask this question for medication noncompliance]

How often do you NOT take your prescription medication as told to by your doctor, nurse or pharmacist?
( ) Most of the time [1]
( ) Some of the time [2]
( ) Seldom or rarely [3]
( ) Never [4]
( ) I don’t know [5]</td></tr>
<tr><td>IF DIAB = yes ask questions TESTDIAB [1, 2, 3, 4 or 5]

and

DIABCON [1, 2, 3, 4 or 5]</td><td>When was the last time your healthcare provider had a hemoglobin A1c blood test done to test your blood sugar?
( ) Within the past 6 months [1]
( ) Between 6 to 12 months ago [2]
( ) Over a year ago [3]
( ) Never [4]
( ) Don’t know [5]

How often is your blood sugar in a healthy range or in the range recommended by your healthcare provider?
( ) Most of the time [1]</td></tr>
</table>

Function/ Question	Question
	( ) Some of the time [2] ( ) Seldom or rarely [3] ( ) Never [4] ( ) I don't know [5]
If DIAB = no ask question SCRNDIAB [1, 2, 3, 4, or 5]	When was the last time the blood sugar level in your blood was tested? ( ) Within the past 12 months [1] ( ) Between 1 to 2 years ago [2] ( ) More than 2 years ago [3] ( ) Never [4] ( ) I don't know [5]
IF HBP = yes then ask TESTHBP [1,2, 3, 4 or 5] and HBPCON [1, 2, 3, 4 or 5] and BPCHANGE (below)	When was the last time your blood pressure was measured by your healthcare provider? ( ) Within the past 6 months [1] ( ) Between 6 to 12 months ago [2] ( ) Over a year ago [3] ( ) Never [4] ( ) Don't know [5] How often is your blood pressure in a healthy range or the range recommended by your healthcare provider? ( ) Most of the time [1] ( ) Some of the time [2] ( ) Seldom or rarely [3] ( ) Never [4] ( ) I don't know [5]
If HBP = no ask question SCRNHBP [1, 2, 3, 4, or 5]	When was the last time your blood pressure was measured by a healthcare provider (e.g., doctor or nurse)? ( ) Within the past 12 months [1] ( ) Between 1 to 2 years ago [2] ( ) More than 2 years ago [3] ( ) Never [4] ( ) I don't know [5]
Readiness to change BP practices (BPCHANGE) [1, 2, 3 or 4]	[If report HBP] You report that your blood pressure may not be in a healthy range most of the time. When would you be willing to start making changes to better manage your high blood pressure? ( ) I'm already trying to make changes to better manage my high blood pressure [1] ( ) I'm willing to start making changes within the next 30 days [2] ( ) I'm willing to start making changes within the next 6 months [3] ( ) I'm not planning on making any changes [4]
IF DYSL = yes then ask TESTDYSL [1,2, 3, 4 or 5] and DYSLCON [1, 2, 3, 4 or 5]	When was the last time your healthcare provider had your blood tested for cholesterol or triglycerides? ( ) Within the past 6 months [1] ( ) Between 6 to 12 months ago [2] ( ) Over a year ago [3] ( ) Never [4] ( ) Don't know [5] How often are your cholesterol or triglyceride levels in healthy ranges or the ranges recommended by your healthcare provider? ( ) Most of the time [1]

Function/ Question	Question
	( ) Some of the time [2] ( ) Seldom or rarely [3] ( ) Never [4] ( ) I don't know [5]
If DYSL = no ask question SCRNDYSL [1, 2, 3, 4, or 5]	When is the last time you had a blood test to measure your blood cholesterol or triglycerides? ( ) Within the past 12 months [1] ( ) Between 1 to 2 years ago [2] ( ) More than 2 years ago [3] ( ) Never [4] ( ) I don't know [5]
Healthcare provider (HCP)	Do you have a healthcare professional, such as a doctor or a nurse practitioner, that you consider your "family doctor "or primary healthcare provider? ( ) Yes [1] ( ) No [2] ( ) Don't know [3]
Where receive healthcare HCLOC1	Where do you go for MOST of your medical care? Please choose only one answer. ( ) The office of my personal physician or nurse practitioner [1] ( ) Walk-in clinics [2] ( ) Hospital emergency departments (ERs) [3] ( ) Other, please specify: [TEXT BOX] [4] [HCLOC2 text box]
Location FAS [text] PROV [	Please give us the first three digits of your postal code. This information will help the Heart and Stroke Foundation in planning programs across the country. _ _ _ [FAS text box] ( ) I don't know my postal code but I live in the province/territory of: [drop-down box of provinces and territories] [PROV: 1=Nfld, 2=NB,3=PEI, 4=NX, 5=QUE, 6=ON, 7=MAN, 8=SASK, 9= ALB, 10=BC, 11=NWT, 12=NUN, 13=YUKON] ( ) I don't live in Canada [XCAN = 1]
Source	How did you learn about this web site? Please choose *all* that apply. ( ) At a doctor's office [MDOFF = 1] ( ) At a pharmacy [PHARM = 1] ( ) A print advertisement in a newspaper [PRINT=1] ( ) A TV advertisement [TV = 1] ( ) Through an Internet search engine, such as Google [SEARCH=1] ( ) An online (Internet) advertisement [DIGIAD=1] ( ) I found it while visiting the Heart & Stroke website [HSF=1] ( ) I found it on another web site [WEB=1] ( ) I received information in the mail [MAIL=1] ( ) I heard about if from a friend, relative, neighbour or co-worker [FRIEND=1] ( ) Other, please specify: [TEXT BOX] [HROTHER=1] [HROTHTXT = text box]
Completed assessment for	I answered these questions: ( ) For myself [1] ( ) For someone else [2] ( ) To investigate or review the site [3]
SES INTRO – If answered questions for self	The Heart and Stroke Foundation is continually working to improve the Health Assessment and the report you receive. To do this, it is helpful to be able to divide responses into large groups. Please help us by answering the following short questions.
MARTIAL	What is your current marital status? Please check one of the following:

Function/ Question	Question
	• I'm married [1] • I have a common-law spouse or partner [2] • I'm widowed, separated or divorced [3] • I'm single and have never been married [4] • I'd rather not answer [5]
EDUCATION	Please pick the one response that best describes the highest level of education you completed, either in Canada or elsewhere in the world. • I didn't complete secondary school (high school) [1] • I completed secondary school high school or CEPEG) [2] • I attended college or university but did not graduate or get a degree or diploma [3] • I compete a college or university program (post-secondary education) and received a degree or diploma [4] • I'd rather not answer [5]
EMPLOYMENT	Which one of the following best describes your current employment status? • I work for wages/have a full-time or part-time job [1] • I have my own business /I'm self-employed [2] • I'm a full-time student [3] • I'm a stay-at-home parent [4] • I'm retired [5] • I'm not able to work for wages [6] • I've been out of work for less than one year [7] • I've been out of work for more than one year [8] • I'd rather not answer [9]
OCCUPATION Ask only if EMPLOYMENT = 1, 2, 7 or 8	Of the following three groups, please select the one that best describes the sort of work you do. • My work is in management, business, finance, administration, natural or applied sciences, health or medicine, social sciences, education, religion, art, culture, or recreation • I work in sales (wholesale or retail) or services (the hospitality industry or personal services such as hairdressers) • I work in the trades (e.g., electrician, plumber, carpenter, mechanic) or hands-on work in the construction or transport industry, as a heavy equipment operator, or in primary industries such as mining, lumbering, fishing, faring, ranching, processing, manufacturing and utilities • I'd rather not say
SES CLOSE	Thank you for answering these questions.
Research consent button CONSENT [1 or 2]	(For those who state they completed risk assessment for themselves) The personal data and contact information you enter in this site will always remain strictly confidential. To help the Heart and Stroke Foundation improve the site, better meet the needs of all users, learn more about the health needs of Canadians and share our learnings through publications in scientific journals, we'd like to anonymously analyze data submitted to this website. Information is "de-identified," meaning that anything that could identify you, such as your email address, is removed. All records are aggregated into one large, anonymous database and analyzed in groups (e.g., looking at the health needs of men compared to women). **Whether or not you agree, you will have free and full access to all of the Heart and Stroke Foundation website services and programs.** Do we have

Function/ Question	Question
	your permission to include your data in our research database? ( ) YES, you can include my data in your anonymous research database. I understand that no information that can personally identify me will be included in the research database. [1] ( ) No, I don't want my information added to the anonymous research database. [2]

Appendix 3: Tables for Chapter 5

Table 1: HRA demographics by gender

	Males (n=38,160)		Females (n=81,900)		Both (n=120,510)		Effect Size
	n	%	n	%	n	%	Cohen's d
Mean (sd) age in yrs	50.02 (14.46)		47.89 (13.94)		48.57 (14.14)		.150
Age group:							**Cramer's V**
18-34	6512	16.9	16540	20.2	23052	19.1	.084
35-44	6843	17.7	14915	18.2	21758	18.1	
45-54	9389	24.3	21837	26.7	31226	25.9	
55-64	9560	24.8	19538	23.9	29098	24.1	
65-74	4940	12.8	7523	9.2	12463	10.3	
75-90	1366	3.5	1547	1.9	2913	2.4	
Ethnicity:							
South Asian	1774	4.6	1372	1.7	3146	2.6	.116
Aboriginal	446	1.2	1432	1.7	1878	1.6	
African/Caribbean	492	1.3	1088	1.3	1580	1.3	
SE Asian	138	0.4	201	0.2	339	0.3	
Other	1622	4.2	3609	4.4	5231	4.3	
Chinese	1815	4.7	2156	2.6	3971	3.3	
Filipino	323	0.8	594	0.7	917	0.8	
Japanese	130	0.3	217	0.3	347	0.3	
Korean	88	0.2	95	0.1	183	0.2	
Arabic	386	1.0	341	0.4	727	0.6	
West Asian	143	0.4	172	0.2	315	0.3	
Latin American	498	1.3	717	0.9	1215	1.0	
White/Caucasian	30724	79.6	69843	85.3	100567	83.5	
Marital status:							
Married	24344	63.1	45854	56.0	70198	58.3	.090
Common-law	4488	11.6	10534	12.9	15022	12.5	
Widow/separated/ divorce	3515	9.1	12315	15.0	15830	13.1	
Single/never	5793	15.0	11937	14.6	17730	14.7	
No response	449	1.2	1221	1.5	1670	1.4	
Highest education:							
< High school	2022	5.2	3255	4.0	5277	4.4	.038
High school	6647	17.2	15842	19.3	22489	18.7	
Some post-secondary	6157	15.9	12335	15.1	18492	15.3	
University/college graduate	23166	60.0	49093	59.9	72259	60.0	
No response	585	1.5	1293	1.6	1878	1.6	
Employment status:							.157

Full/part-time	21563	55.8	47717	58.3	69280	57.5	
Self-employed	5064	13.1	6124	7.5	11188	9.3	
Full-time student	1318	3.4	3509	4.3	4827	4.0	
Stay-at-home parent	131	0.3	4323	5.3	4454	3.7	
Retired	8133	21.1	14363	17.5	22496	18.7	
Permanently unable to work	551	1.4	1727	2.1	2278	1.9	
Unemployed > 1 yr	696	1.8	1500	1.8	2196	1.8	
Unemployed > 1 yr	633	1.6	1203	1.5	1836	1.5	
No response	501	1.3	1382	1.7	1883	1.6	
Type of work:							
Mgmt, health, education	21690	56.2	55832	68.2	77522	64.3	.278
Sales/service	4223	10.9	9868	12.0	14091	11.7	
Trades	7879	20.4	2882	3.5	10761	8.9	
No response	4509	11.7	12360	15.1	16869	14.0	

All comparisons statistically significant (p<.001) by independent t-test (continuous variables) or Chi square (categorical variables). For Cohen's d, 0.2 = small effect, 0.5 medium effect, and 0.8 large effect. For Cramer's V for 1 degree of freedom (two categories) 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect.

Table 2: Demographic variables by age group

	18-34 yrs (n= 23,052)	**35-44 yrs (n= 21,758)**	**45-54 yrs (n= 31,226)**	**55-64 yrs (n= 29.098)**	**65-74 yrs (n= 12,463)**	**75-90 yrs (n= 2,931)**	**Effect size by age *eta***
% male gender	28.2	31.5	30.1	32.9	39.6	46.9	.084
% higher education	81.8	85.3	76.4	72.8	65.7	56.7	.157
% married/ common-law	31.1	62.3	63.9	66.7	67.5	60.3	.271
% employed full/part-time	66.0	75.9	74.2	48.3	12.1	2.2	.433
% white collar occupation	66.6	71.4	69.5	63.3	48.4	41.1	.156

For eta (effect by interval), 0.01 = small effect, 0.06 = medium-sized effect, and 0.14 = large effect; all comparisons significant at p<.001.

Table 3: Report of non-modifiable risk factors by gender

Non-modifiable risk factor	Males (n=38,160)		Females (n=81,900)		Both (n=120,510)		Effect size
	n	%	n	%	n	%	Cramer's V
% higher risk ethnicity	2712	7.0	3892	4.8	6604	5.5	.047
% family history premature stroke	5437	14.1	13533	16.5	18970	15.7	.039
% family history premature heart disease	12568	32.6	32576	39.8	45144	37.5	.074
% family history dyslipidemia	15984	41.4	37843	46.2	53827	44.7	.045
% family history hypertension	20608	53.4	49318	60.2	69926	58.0	.065
% family history diabetes	15172	39.3	38821	47.4	53993	44.8	.078

All comparisons statistically significant (p<.001) as estimated by Chi squares. For Cramer's V for 1 degree of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect.

Table 4: Prevalence of non-modifiable risk factors by age group

Risk factor/ condition	18-34 yrs (n= 23,052)	35-44 yrs (n= 21,758)	45-54 yrs (n= 31,226)	55-64 yrs (n= 29.098)	65-74 yrs (n= 12,463)	75-90 yrs (n= 2,931)	Effect size by age *eta*
% higher risk ethnicity	8.5	8.2	4.9	3.1	2.8	2.4	.102
% family history stroke	15.7	17.3	16.4	15.6	13.5	12.9	.031
% family history dyslipidemia	47.2	47.3	45.8	44.4	37.4	31.7	.071
% family history diabetes	48.1	47.8	46.1	43.7	37.5	32.6	.074
% family history hypertension	55.7	58.4	60.2	59.5	55.7	52.4	.041
% family history premature heart disease	31.8	38.4	40.2	39.8	36.4	32.8	.032

For eta (effect by interval), 0.01 = small effect, 0.06 = medium-sized effect, and 0.14 = large effect; all comparisons significant at p<.001.

Table 5: Modifiable risk factors and associated stage of change by gender

Risk Factor or Stage	Males (n=38,610) n	%	Females (n=81,900) n	%	Both (n=120,510) n	%	Effect size Cramer's V
Physical inactivity	16657	43.1	42902	52.4	59559	49.4	.091
Stage of change for those inactive:							
Precontemplation	2510	15.1	5873	13.7	8383	14.1	
Contemplation	5796	34.8	17249	40.2	23045	38.7	.061
Preparation	6023	36.2	15197	35.4	21220	35.6	
Action	2328	14.0	4583	10.7	3944	11.6	
Unwilling (top 2)	8306	49.9	23122	53.9	31428	52.8	.036
Willing (bottom 2)	8351	50.1	19781	46.1	28131	47.2	
Smoking	4965	13.0	10115	12.4	15080	12.5	.007 (p=.013)
Stage for smokers:							
Precontemplation	1119	22.5	2499	24.7	3618	24.0	
Contemplation	682	13.7	1406	13.9	2088	13.8	.031
Preparation	1932	38.9	3937	38.9	5869	38.9	(p=.002)
Action	1232	24.8	2273	22.5	3505	23.2	
Unwilling (top 2)	1801	36.3	3905	38.6	5706	37.8	.023
Willing (bottom 2)	3164	63.7	6210	61.4	9374	62.2	(p=.006)
Overweight/obese or at-risk waist circumference	25300	65.5	49225	60.1	74525	61.8	.052
Stage of change for overweight/obese:							
Precontemplation	3252	12.9	4471	9.1	7723	10.4	
Contemplation	7996	31.6	15175	30.8	23171	31.1	.063
Preparation	10772	42.6	22714	46.1	33486	44.9	
Action	3280	13.0	6865	13.9	10145	13.6	
Unwilling (top 2)	11248	44.5	19646	39.9	30894	41.5	.044
Willing (bottom 2)	14052	55.5	29579	60.1	43631	58.5	
Excess alcohol	12294	31.8	16764	20.5	29058	24.1	.124
Stage of change for unsafe drinkers:							
Precontemplation	2329	18.9	2654	15.8	4983	17.1	
Contemplation	3176	25.8	4983	29.7	8159	28.1	.084
Preparation	3514	28.6	5544	33.1	9058	31.2	
Action	3275	26.6	3583	21.4	6858	23.6	
Unwilling (top 2)	5505	44.8	7637	45.6	13142	45.2	.008
Willing (bottom 2)	6789	55.2	9127	54.4	15916	54.8	(p=.188)
Fatty foods frequency:							
< 1/week	20005	51.8	49797	60.8	69802	57.9	
2-3 times/week	12716	32.9	21812	26.6	34528	28.7	.037
3+ times/week	5725	14.8	9986	12.2	15711	13.0	
Fast foods frequency:							
< 1/week	27311	70.7	66348	81.0	93659	77.7	
2-3 times/week	9462	24.5	13359	16.3	22821	18.9	.057

Risk Factor or Stage	Males (n=38,610)		Females (n=81,900)		Both (n=120,510)		Effect size Cramer's V
3+ times/week	1631	4.2	1809	2.2	3440	2.9	
Fish consumption frequency:							
< 1/week	19468	50.4	44372	54.2	63840	53.0	.035
2-3 times/week	15342	39.7	31507	38.5	46849	38.9	
3+ times/week	3616	9.4	5692	6.9	9308	7.7	
5 servings fruit/vegetables/day:							
< 1/week	7152	18.5	10623	13.0	17775	14.7	.107
2-3 times/week	12041	31.2	20872	25.5	32913	27.3	
3+ times/week	19339	50.1	50267	61.4	69606	57.8	
≥1 bad diet behaviour	27911	72.3	55902	68.3	83813	69.5	.041
Stage of change for those with poor diet:							
Precontemplation	2679	9.6	3163	5.7	5842	7.0	.093
Contemplation	5311	19.0	9719	17.4	15030	17.9	
Preparation	13696	49.1	31992	57.2	45688	54.5	
Action	6225	22.3	11028	19.7	17253	20.6	
Unwilling (top 2)	7990	28.6	12882	23.0	20872	24.9	.061
Willing (bottom 2)	19921	71.4	43020	77.0	62941	75.1	
Stress frequency:							
Seldom/never	14759	38.2	20158	24.6	34917	29.0	.105
Few times	18527	48.0	43122	52.7	61649	51.2	
Often/most	5270	13.6	18512	22.6	23782	19.7	
***Stage of change for frequently stressed**:*							
Precontemplation	560	10.5	1191	6.4	1751	7.3	.074
Contemplation	1307	24.5	4669	25.1	5976	24.9	
Preparation	2369	44.4	9321	50.0	11690	48.8	
Action	1095	20.5	3448	18.5	4543	19.0	
Unwilling (top 2)	1867	35.0	5860	31.5	7727	32.2	.032
Willing (bottom 2)	3464	65.0	12769	68.5	16233	67.8	
Salt consumption:							
Try to reduce salt	19418	50.3	45392	55.4	64810	53.8	.055
Don't monitor salt	16667	43.2	8064	9.8	24731	20.5	
Eat a lot of salty foods	14562	37.7	28310	34.6	42872	35.6	
Stage of change for salt:							
Precontemplation	2007	16.7	2746	13.0	4753	14.3	.077
Contemplation	1516	12.6	2345	11.1	3861	11.6	
Preparation	4075	33.9	8677	41.0	12752	38.5	
Action	4417	36.8	7373	34.9	11790	35.6	
Unwilling (top 2)	3523	29.3	5091	24.1	8614	26.0	.057
Willing (bottom 2)	8492	70.7	16050	75.9	24542	74.0	

Except where indicated, comparisons were statistically significant (p<.001) as estimated by Chi squares. For Cramer's V for 1 degree of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect.

Table 6: Prevalence of modifiable risk factors by age group

Modifiable risk factor	18-34 yrs (n= 23,052)	35-44 yrs (n= 21,758)	45-54 yrs (n= 31,226)	55-64 yrs (n= 29,098)	65-74 yrs (n= 12,463)	75-90 yrs (n= 2,931)	Effect size by age *eta*
% physical inactivity	47.0	50.4	45.1	40.0	33.9	30.4	.095
% smoking	15.5	14.7	14.3	10.5	6.1	2.6	.091
% excess alcohol	25.4	22.0	24.7	24.6	23.8	20.3	.006
% frequent stress	25.2	26.0	21.7	15.0	8.0	6.7	.144
% over-weight/ obese	51.3	63.2	65.0	65.8	63.2	55.9	.070
% high salt	43.5	34.2	25.2	18.9	15.4	13.8	.214
% high fat foods	20.3	15.9	12.1	9.3	7.5	6.8	.128
% fast foods	6.1	4.0	2.4	1.2	0.6	0.6	.110
% low fruit/veg	50.1	47.5	41.9	36.3	34.0	34.1	.115
% low fish	58.5	60.0	55.8	47.3	42.0	38.9	.120
% $\geq$1 bad dietary behaviour	76.9	75.7	71.1	63.4	58.8	56.3	.139

For eta (effect by interval), 0.01 = small effect, 0.06 = medium-sized effect, and 0.14 = large effect; all comparisons significant at p<.001 unless indicated otherwise.

Table 7: Of those with modifiable risk factor, readiness to change by age group

Modifiable risk factor	18-34 yrs (%)	35-44 yrs (%)	45-54 yrs (%)	55-64 yrs (%)	65-74 yrs (%)	75-90 yrs (%)	All ages (%)	Effect size *eta*
% Physical activity (n=59,559)	14.4	14.3	13.4	14.0	14.5	16.5	14.1	.048
Precontemplation	38.5	41.2	39.2	37.7	36.1	29.9	38.7	
Contemplation	38.6	34.4	35.8	36.3	35.4	30.1	35.6	
Preparation	10.4	10.2	11.6	12.1	14.0	23.5	11.6	
Action								
Unwilling (top 2)	52.9	55.5	52.6	51.7	50.6	46.4	52.8	.024
Willing (bottom 2)	47.1	44.5	47.4	48.3	49.4	53.6	47.2	
% Smoking (n=15,080)								
Precontemplation	24.4	25.6	25.2	21.3	18.6	26.3	24.0	.038
Contemplation	14.0	13.9	14.8	13.4	9.1	15.8	13.8	
Preparation	39.0	38.5	36.4	41.3	46.6	30.3	38.9	
Action	22.6	22.0	23.6	24.1	25.7	27.6	23.2	
Unwilling (top 2)	38.4	39.5	40.1	34.6	27.7	42.1	37.8	.038
Willing (bottom 2)	61.6	60.5	59.9	65.4	72.3	57.9	62.2	
% Alcohol (n=29,058)								
Precontemplation	16.3	17.0	17.8	17.2	17.3	16.9	17.1	.088
Contemplation	20.6	26.2	30.3	32.8	30.1	22.0	28.1	
Preparation	36.8	33.4	30.2	27.9	28.0	25.1	31.2	
Action	26.3	23.5	21.7	22.1	24.6	35.9	23.6	
Unwilling (top 2)	36.9	43.1	48.1	50.0	47.4	39.0	45.2	.074
Willing (bottom 2)	63.1	56.9	51.9	50.0	52.6	61.0	54.8	
% Weight (n=74,525)								
Precontemplation	12.6	11.7	9.9	9.0	9.3	9.2	10.4	.047
Contemplation	30.5	32.0	31.3	30.8	31.2	27.9	31.1	
Preparation	42.7	43.1	44.5	46.7	47.7	47.6	44.9	
Action	14.1	13.1	14.3	13.5	11.9	15.3	13.6	
Unwilling (top 2)	43.2	43.7	41.2	39.8	40.5	37.1	41.5	.029
Willing (bottom 2)	56.8	56.3	58.8	60.2	59.5	62.9	58.5	
% Diet (n=83,813)	9.1	7.8	6.7	5.6	4.8	3.9	7.0	.085
Precontemplation	20.3	19.9	17.6	16.2	14.6	12.0	17.9	
Contemplation	51.9	53.4	54.1	56.4	59.3	56.3	54.5	
Preparation	18.7	18.9	21.6	21.8	21.3	27.7	20.6	
Action								
Unwilling (top 2)	29.4	27.7	24.3	21.8	19.4	15.9	24.9	.081
Willing (bottom 2)	70.6	72.3	75.7	78.2	80.6	84.1	75.1	
% Stress (n=23,960)								
Precontemplation	9.7	8.1	6.4	5.2	4.5	5.3	7.3	.104
Contemplation	28.2	27.0	24.0	20.9	20.5	16.0	24.9	
Preparation	45.4	47.3	50.0	52.5	52.2	47.6	48.8	
Action	16.7	17.5	19.5	21.4	22.8	31.1	19.0	
Unwilling (top 2)	37.9	35.2	30.4	26.0	25.0	21.4	32.2	.097
Willing (bottom 2)	62.1	64.8	69.6	74.0	75.0	78.6	67.8	

Modifiable risk factor	18-34 yrs (%)	35-44 yrs (%)	45-54 yrs (%)	55-64 yrs (%)	65-74 yrs (%)	75-90 yrs (%)	All ages (%)	Effect size *eta*
% Salt (n=33,156)	17.4	14.1	12.8	12.6	10.5	13.5	14.3	.115
Precontemplation	14.3	12.0	10.9	9.0	7.9	8.0	11.6	
Contemplation	38.0	40.4	39.4	37.0	36.8	26.2	38.5	
Preparation	30.4	33.5	36.8	41.5	44.8	52.4	35.6	
Action								
Unwilling (top 2)	31.7	226.1	23.8	21.5	18.4	21.4	26.0	.093
Willing (bottom 2)	68.3	73.9	76.2	78.5	81.6	78.6	74.0	

For eta (effect by interval), 0.01 = small effect, 0.06 = medium-sized effect, and 0.14 = large effect; all comparisons significant at p<.001 unless indicated otherwise.

Table 8: Prevalence of select chronic diseases by gender

Medical Diagnoses	Males (n=38,610)		Females (n=81,900)		Both (n=120,510)		Effect size Cramer's V
	n	%	n	%	n	%	
Diabetes	3649	9.5	4592	5.6	8241	6.8	.071
Heart disease	2983	7.7	2372	2.9	5355	4.4	.109
Hypertension	12545	32.5	18955	23.1	31500	26.1	.099
Dyslipidemia	10775	27.9	14301	17.5	25076	20.8	.120
Stroke	1045	2.7	1465	1.8	2510	2.1	.030
Alzheimers	266	0.7	302	0.4	568	0.5	.022
Arthritis	4908	12.7	15566	19.0	20474	17.0	.078
Asthma	2792	7.2	8948	10.9	11740	9.7	.058
Cancer	1452	3.8	2746	3.4	4198	3.5	.010
COPD	833	2.2	1543	1.9	2376	2.0	.009
Back pain	3320	8.6	7628	9.3	10948	9.1	.012
Mood disorders	4349	11.3	16066	19.6	20415	16.9	.104
Renal disease	659	1.7	876	1.1	1535	1.3	.027
Liver disease	568	1.5	822	1.0	1390	1.2	.020
Osteoporosis	757	2.0	5973	7.3	6730	5.6	.108
Sleep apnea	3495	9.1	3421	4.2	6916	5.7	.098

All comparisons were statistically significant ($p<.001$) as estimated by Chi squares. For Cramer's V for 1 degree of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect.

Table 9: Prevalence of chronic diseases by age group

Risk factor/ condition	18-34 yrs (n= 23,052)	35-44 yrs (n= 21,758)	45-54 yrs (n= 31,226)	55-64 yrs (n= 29.098)	65-74 yrs (n= 12,463)	75-90 yrs (n= 2,931)	Effect size by age *eta*
% diabetes	1.9	3.8	6.5	10.0	13.1	14.5	.148
% heart disease	1.0	1.5	3.3	6.1	11.4	19.6	.172
% hypertension	5.9	14.5	25.7	37.7	50.0	59.4	.338
% dyslipidemia	4.8	11.5	20.0	31.9	38.8	37.6	.282
% stroke	0.8	1.2	1.7	2.6	4.2	8.8	.087
% Alzheimers	0.6	0.5	0.4	0.4	0.5	1.1	.019
% arthritis	3.5	7.6	15.5	26.8	33.9	40.5	.289
% asthma	12.4	10.7	9.5	8.1	8.0	6.8	.056
% cancer	1.1	1.8	3.0	4.8	7.5	10.1	.119
% COPD	0.8	0.9	1.6	2.4	4.7	6.2	.095
% back pain	5.4	8.0	9.5	10.8	11.4	15.3	.078
% mood disorder	16.9	19.4	18.3	16.6	12.0	8.9	.063
% renal disease	0.9	1.0	1.2	1.4	1.9	3.0	.035
% liver disease	0.9	1.9	1.3	1.4	1.0	0.8	.018
% osteoporosis	0.9	1.4	3.6	9.6	14.1	19.7	.216
% sleep apnea	2.1	4.4	6.3	8.0	7.9	6.8	.094
% prescribed medication	22.0	31.2	41.1	54.9	66.1	74.2	.301
% some/most time miss medication	22.1	17.3	13.9	9.6	6.7	5.6	.143

For eta (effect by interval), 0.01 = small effect, 0.06 = medium-sized effect, and 0.14 = large effect; all comparisons significant at p<.001.

Table 10: Total number of modifiable and medical CVD risk factors by gender

Total number of modifiable & medical risk factors	**Males (n=38,610)**		**Females (n-81,900)**		**Both (n=120,510)**		**Effect size Cohen's d**
Mean (sd)	3.34 (1.65)		3.03 (1.62)		3.13 (1.64)		.190
Number:	**n**	**%**	**n**	**%**	**n**	**%**	**Cramer's V**
0	1127	2.9	3767	4.6	4894	4.1	.091
1	3903	10.1	10998	13.5	14901	12.4	
2	7346	19.1	17333	21.2	24679	20.5	
3	9082	23.6	19435	23.8	28517	23.7	
4	8028	20.8	15338	18.8	23366	19.4	
5	5114	13.3	9114	11.2	14228	11.8	
6	2638	6.8	4059	5.0	6697	5.6	
7	956	2.5	1312	1.6	2268	1.9	
8	267	0.7	3	0.3	5	0.5	
9	57	0.1	59	0.1	116	0.1	
10	4	0.0	7	0.0	11	90.0	
Missing:	88	0.2	201	0.2	289	0.2	

All comparisons statistically significant (p<.001).

For Cohen's d, 0.2 = small effect, 0.5 medium effect, and 0.8 large effect. For Cramer's V for 1 degrees of freedom (two categories) 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect.

Table 11: Means by age group

Number of:	Total	18-34 yrs	35-44 yrs	45-54 yrs	55-64 yrs	65-74 yrs	75-90 yrs	Effect size
	x **(sd)**	***x*** **(sd)**	***x*** **(sd)**	***x*** **(sd)**	***x*** **(sd)**	***x*** **(sd)**	***x*** **(sd)**	**ϖ**
Modifiable risk factors	2.6 (1.4)	2.9 (1.4)	2.9 (1.4)	2.7 (1.4)	2.4 (1.3)	2.1 (1.2)	1.9 (1.2)	.200
Non-modifiable risk factors	2.1 (1.5)	2.1 (1.5)	2.2 (1.5)	2.1 (1.4)	2.1 (1.4)	1.8 (1.4)	1.6 (1.3)	.079
Vascular conditions	0.6 (1.0)	0.2 (0.6)	0.3 (0.7)	0.6 (0.9)	0.9 (1.0)	1.2 (1.1)	1.4 (1.2)	.369
Total CVD risk factors	3.1 (1.6)	3.0 (1.5)	3.2 (1.6)	3.2 (1.7)	3.2 (1.7)	3.1 (1.6)	3.0 (1.5)	.051

All comparison statistically significant (p<.001) as estimated by ANOVA. For ϖ standard cut-offs are: 0.01 = small effect, 0.06 = medium effect and 0.14 = large effect.

Table 12: Hypertension screening and management by gender

	Males (n=38,610)		Females (n-81,900)		Both (n=120,510)		Effect size
	n	**% of males**	**n**	**% of females**	**n**	**% of both**	**Cramer's V**
Normotensive	26065	67.5%	62945	76.9%	89010	73.9%	0.99
Screening of blood pressure of normotensives:							
< 12 mos	20590	75.0%	52390	81.0%	72980	79.2%	.084
1-2 yrs	3447	12.6%	7516	11.6%	10963	11.9%	
> 2 yrs	2110	7.7%	3129	4.8%	5239	5.7%	
Never	628	2.3%	748	1.2%	1376	1.5%	
Don't know	666	2.4%	917	1.4%	1583	1.7%	
Hypertensive	12545	32.5%	18955	23.1%	31500	26.1%	0.99
Prescribed medication	9233	39.8%	14428	61.0%	23661	75.1%	0.29
Last time BP of hypertensives measured:							
< 6 mos	9598	76.5%	14877	78.5%	24475	77.7%	.032
6-12 mos	904	7.2%	1413	7.5%	2317	7.4%	
> 1 yr	458	3.7%	521	2.7%	979	3.1%	
Never	45	0.4%	44	0.2%	89	0.3%	
Don't know	51	0.4%	59	0.3%	110	0.3%	
Blood pressure controlled (in healthy range):							
Most of time	6534	52.1%	10553	55.7%	17087	54.2%	.049
Some of time	2729	21.8%	4204	22.2%	6933	22.0%	
Seldom/rarely	1163	9.3%	1418	7.5%	2581	8.2%	
Never	352	2.8%	399	2.1%	751	2.4%	
Don't know	293	2.3%	374	2.0%	667	2.1%	

All comparisons were statistically significant (p<.001) by independent t-test (continuous variable) or Chi squares (categorical variables). For Cramer's V for 1 degree of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect.

Table 13: Hypertension management by age group

	18-34 yrs (%)	**35-44 yrs (%)**	**45-54 yrs (%)**	**55-64 yrs (%)**	**65-74 yrs (%)**	**75-90 yrs (%)**	**Effect size *eta***
% normotensive	94.1	85.5	74.3	62.3	50.0	40.6	.338
Screening of blood pressure of normotensives (%):							
< 12 mos	69.3	73.9	81.5	87.2	91.6	93.9	
1-2 yrs	14.2	15.4	11.9	8.6	5.8	4.0	.211
> 2 yrs	7.7	7.9	5.3	3.4	2.0	1.6	
Never	4.3	1.2	0.5	0.3	0.2	0.3	
Don't know	4.5	1.6	0.7	0.4	0.4	0.2	
% hypertensive	5.9	14.5	25.7	37.7	50.0	59.4	.338
% prescribed medication	39.0	62.6	71.7	79.2	83.9	84.7	.214
Last time BP of hypertensive measured (%):							
< 6 mos	73.4	82.8	86.8	89.1	90.3	91.7	
6-12 mos	13.4	10.1	8.2	7.8	7.6	7.1	.127
> 1 yr	9.2	6.1	4.4	2.8	1.6	0.9	
Never	1.9	0.5	0.3	0.2	0.2	0.1	
Don't know	2.1	0.6	0.3	0.2	0.4	0.2	
Blood pressure controlled (in healthy range) (%):							
Most of time	34.9	43.4	53.3	66.6	72.6	74.7	
Some of time	31.9	31.0	28.2	22.7	20.4	19.4	.246
Seldom/rarely	19.0	15.9	12.0	7.1	4.9	4.3	
Never	8.7	5.0	3.3	1.8	1.4	0.7	
Don’t know	5.5	4.7	3.2	1.8	0.7	1.0	
Stage of change for uncontrolled hypertensives (%):							
Action	61.9	67.3	71.8	76.3	78.5	75.0	
Preparation	29.1	27.8	23.7	20.1	18.5	21.0	.105
Contemplation	6.8	4.0	3.6	2.5	1.9	1.9	
Precontemplation	2.2	1.0	0.9	1.0	1.1	2.2	

For eta (effect by interval), 0.01 = small effect, 0.06 = medium-sized effect, and 0.14 = large effect; all comparisons significant at p<.001.

Table 14: Dyslipidemia screening and management by gender

	Males (n=38,610) n %		Females (n-81,900) n %		Both (n=120,510) n %		Effect size Cramer's V
No diagnosis of dyslipidemia	27835	72.1	67599	82.5	95434	79.2	.120
For those without dyslipidemia, last time lipids tested:							
< 12 mos	16042	55.2	37566	54.8	53608	54.9	.034
1-2 yrs	4419	15.2	11831	17.2	16250	16.6	
> 2 yrs	3126	10.8	6451	9.4	9577	9.8	
Never	3027	10.4	6568	9.6	9595	9.8	
Don't know	2425	8.4	6183	9.0	8608	8.8	
Report dyslipidemia	10775	27.9	14301	17.5	25076	20.8	.120
Prescribed medication	7334	68.1	8254	57.7	15588	62.2	.106
For dyslipidemics, last time lipids tested:							
< 6 mos	6428	68.2	8315	64.2	14743	65.9	.042
6-12 mos	1829	19.5	2886	22.3	4715	21.1	
> 1 yr	1061	11.3	1595	12.3	2656	11.9	
Never	44	0.5	61	0.5	105	0.5	
Don't know	63	0.7	92	0.7	155	0.7	
Lipids controlled (in a healthy range):							
Most of time	5067	47.0	5870	41.0	10937	43.6	.087
Some of time	2272	21.1	3438	24.0	5710	22.8	
Seldom/rarely	987	9.2	1658	11.6	2645	10.5	
Never	624	5.8	1137	8.0	1761	7.0	
Don't know	486	4.5	850	5.9	1336	5.3	

All comparisons were statistically significant (p<.001) by Chi squares. For Cramer's V for 1 degree of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect.

Table 15: Dyslipidemia management by age group

	18-34 yrs (%)	**35-44 yrs (%)**	**45-54 yrs (%)**	**55-64 yrs (%)**	**65-74 yrs (%)**	**75-90 yrs (%)**	**Effect size *eta***
% normal lipids	95.2	88.5	80.0	68.1	61.2	62.4	.282
For those without dyslipidemia, last time lipids tested (%):							
< 12 mos	29.5	45.2	60.6	71.3	80.3	80.8	
1-2 yrs	13.8	19.5	19.4	15.9	12.0	9.9	.413
> 2 yrs	9.8	13.8	10.6	7.7	4.7	4.2	
Never	25.7	11.2	4.6	2.4	1.2	2.1	
Don't know	21.0	10.3	4.8	2.7	1.9	3.0	
% dyslipidemic	4.8	11.5	20.0	31.9	36.8	37.6	.282
% prescribed medication	18.3	36.6	55.4	68.6	77.5	81.5	.309
For dyslipidemics, last time lipids tested (%):							
< 6 mos	43.0	57.0	63.9	69.1	72.1	69.5	
6-12 mos	20.2	20.4	20.5	21.1	21.6	24.5	.211
> 1 yr	28.6	20.3	14.6	9.4	5.8	5.1	
Never	3.4	1.0	0.4	0.1	0.2	0.4	
Don't know	4.8	1.2	0.6	0.3	0.3	0.4	
Lipids controlled (in a healthy range) (%):							
Most of time	25.2	28.5	41.0	52.8	64.5	68.7	
Some of time	25.4	28.8	28.9	25.7	20.8	16.2	.265
Seldom/rarely	15.3	17.9	14.7	10.6	7.3	6.1	
Never	19.3	14.1	8.6	6.1	4.6	5.5	
Don’t know	14.8	10.7	6.8	4.8	2.9	3.5	

For eta (effect by interval), 0.01 = small effect, 0.06 = medium-sized effect, and 0.14 = large effect; all comparisons significant at p<.001.

Table 16: Diabetes screening and management by gender

	Males (n=38,610)		Females (n-81,900)		Both (n=120,510)		Effect size
	n	%	n	%	n	%	Cramer's V
No diagnosis of diabetes	34961	90.5	77308	94.4	112269	93.2	.071
For those without diabetes, last time blood glucose tested:							
< 12 mos	20624	58.3	45490	58.6	66114	58.5	.050
1-2 yrs	4693	13.3	11910	15.3	16603	14.7	
> 2 yrs	3099	8.8	7572	9.8	10671	9.4	
Never	3743	10.6	6267	8.1	10010	8.9	
Don't know	3191	9.0	6404	8.2	9595	8.5	
Report diabetes	3649	9.5	4592	5.6	8241	6.8	.071
Prescribed medication	2622	71.8	3100	67.5	5722	69.4	.047
For those with diabetes, last time glucose tested (A1c):							
< 6 mos	2295	62.9	2883	62.8	5178	62.8	.032
6-12 mos	365	10.0	477	10.4	842	10.2	(p=.108)
> 1 yr	148	4.1	195	4.2	343	4.2	
Never	77	2.1	93	2.0	170	2.1	
Don't know	332	9.1	512	11.1	844	10.2	
Glucose Control:							
Most of time	2029	55.6	2480	54.0	4509	54.7	.043
Some of time	769	21.1	1113	24.2	1882	36.3	(p=.010)
Seldom/rarely	224	6.1	284	6.2	508	9.8	
Never	107	2.9	175	3.8	282	5.4	
Don't know	81	2.2	97	2.1	178	3.4	

Except when indicated otherwise, Chi squares showed comparisons were statistically significant (p<.001). For Cramer's V for 1 degree of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect.

Table 17: Diabetes management by age group

	18-34 yrs (%)	**35-44 yrs (%)**	**45-54 yrs (%)**	**55-64 yrs (%)**	**65-74 yrs (%)**	**75-90 yrs (%)**	**Effect size *eta***
% no diabetes	98.1	96.2	93.5	90.0	86.9	85.5	.148
Of those without diabetes, last time blood glucose tested (%):							
< 12 mos	35.8	48.1	61.4	72.2	79.9	79.7	
1-2 yrs	14.3	17.5	16.5	13.5	9.6	8.0	.333
> 2 yrs	11.4	14.2	9.9	6.3	3.9	3.6	
Never	21.6	10.3	5.6	3.4	2.6	4.2	
Don't know	16.9	9.9	6.6	4.5	4.0	4.6	
% diabetes	1.9	3.8	6.5	10.0	13.1	14.5	.148
% prescribed medication	54.3	64.2	69.4	71.8	71.7	71.2	.074
Of those with diabetes, last time blood glucose tested (A1c) (%):							
< 6 mos	51.4	58.8	68.0	73.7	76.2	75.1	
6-12 mos	12.8	13.7	11.4	11.3	10.4	10.0	.185
> 1 yr	11.1	8.0	5.2	3.4	3.0	3.3	
Never	10.8	5.4	2.5	1.0	0.8	1.1	
Don't know	13.9	14.1	12.9	10.5	9.5	10.5	
Blood glucose controlled (in healthy range) (%):							
Most of time	47.0	53.8	55.4	63.8	67.8	78.1	
Some of time	25.4	25.2	27.6	26.0	24.8	15.8	.186
Seldom/rarely	8.3	11.2	9.0	5.9	4.4	3.1	
Never	13.3	6.6	4.2	2.6	2.1	1.4	
Don’t know	6.0	3.2	3.8	1.6	1.0	1.7	

For eta (effect by interval), 0.01 = small effect, 0.06 = medium-sized effect, and 0.14 = large effect; all comparisons significant at p<.001.

Appendix 4: Detailed Tables for Chapter 6

Table 1: Distribution of general CCHS and HRA populations by age and gender

Age Group	*Canadian Population*		*HRA Population*		*Odds Ratio (95% CI)*	*p*
	n	*%*	*n*	*%*		
Males						
20-34 yrs	3,635,200	27.9	6,166	16.1	0.50 (0.48-0.51)	<.001
35-44 yrs	2,355,200	18.1	6,843	17.9	0.98 (0.96-1.01)	0.19
45-64 yrs	4,877,800	37.5	18,949	49.5	1.64 (1.60-1.67)	<.001
65-89 yrs	2,149,000	16.5	6,306	16.5	0.99 (0.97-1.02)	0.02
Subtotal	13,017,200	100.0	38,264	100.0		
Females						
20-34 yrs	3,534,300	26.4	15,740	19.4	0.68 (0.66-0.68)	<.001
35-44 yrs	2,331,900	17.4	14,915	18.4	1.07 (1.05-1.09)	<.001
45-64 yrs	4,933,500	36.8	41,375	51.0	1.78 (1.76-1.81)	<.001
65-89 yrs	2,588,400	19.3	9,070	11.2	0.53 (0.51-0.54)	<.001
Subtotal	13,388,100	100.0	81,100	100.0		
Both						
20-34 yrs	7,169,600	27.2	21,906	18.4	0.60 (0.59-0.61)	<.001
35-44 yrs	4,687,100	17.8	21,758	18.2	1.03 (1.02-1.05)	<.001
45-64 yrs	9,811,400	37.2	60,324	50.5	1.73 (1.71-1.75)	<.001
65-89 yrs	4,737,400	17.9	15,376	12.9	0.68 (0.66-0.69)	<.001
Total	26,405,500	100.0	118,364	100.0		

Table 2: Highest level of education by age group and gender, CCHS (weighted) and HRA populations (unweighted)

	CCHS Canadian Population		***Unweighted HRA Population (n=86,589)***		***Odds Ratio (95% CI)***
	#	**%**	**#**	**%**	
Males					
< High School					
20-34 yrs	29079	16.7	153	7.8	0.42 (0.36-0.49)
35-44 yrs	194540	11.2	201	10.2	0.81 (0.71-0.94)
45-64 yrs	633937	36.4	934	47.4	1.57 (1.44-1.72)
65-89 yrs	621160	35.7	682	34.6	0.95 (0.87-1.05)*
Subtotal	1739716	100.0	1970	100.0	
High School					
20-34 yrs	569332	29.7	771	12.2	0.33 (0.31-0.35)
35-44 yrs	361876	18.8	750	11.9	0.60 (0.55-0.64)
45-64 yrs	752165	39.2	3439	54.4	1.85 (1.76-1.95)
65-89 yrs	236749	12.3	1364	21.6	1.96 (1.84-2.08)
Subtotal	1920122	100.0	6324	100.0	
Some post-secondary					
25-34 yrs	401792	44.8	937	15.6	0.23 (0.22-0.24)
35-44 yrs	122795	13.7	973	16.2	1.17 (1.09-1.26)
45-64 yrs	263371	29.4	3087	51.3	2.53 (2.41-2.67)
65-89 yrs	107945	12.0	1016	16.9	1.48 (1.39-1.59)
Subtotal	895903	100.0	6013	100.0	
University/College graduate					
25-34 yrs	2004798	26.7	4059	18.0	0.60 (0.58-0.62)
35-44 yrs	1659949	22.1	4622	20.5	0.91 (0.88-0.94)
45-64 yrs	2881497	38.4	10848	48.0	1.49 (1.45-1.53)
65-89 yrs	967338	12.9	3057	13.5	1.06 (1.12-1.10)
Subtotal	7513582	100.0	22586	100.0	
Males – All Ages					
< High School	1739716	23.8	1970	5.3	0.12 (0.11-0.12)
High School	1920211	36.2	6324	17.1	0.36 (0.35-0.37)
Some post-secondary	895903	16.9	6013	16.3	0.96 (0.93-0.99)
Univ/Coll graduate	751382	34.2	22586	61.2	9.58 (9.37-9.78)
Subtotal	5307212	100.0	36893	100.0	
Females					
< High School					
25-34 yrs	203649	11.0	341	10.7	0.97 (0.87-1.08)*
35-44 yrs	135085	7.3	342	10.8	1.53 (1.36-1.71)
45-64 yrs	611565	33.1	1655	52.0	2.19 (2.04-2.35)
65-89 yrs	896795	48.6	846	26.5	0.38 (0.35-0.41)
Subtotal	1847094	100.0	3181	100.0	
High School					
25-34 yrs	469038	22.2	2007	13.2	0.53 (0.51-0.56)
35-44 yrs	332535	15.7	1812	11.9	0.72 (0.69-0.76)
45-64 yrs	893525	42.3	8911	58.4	1.92 (1.86-1.99)
54-89 yrs	419709	19.8	251	16.5	0.80 (0.77-0.83)
Subtotal	2114807	100.0	152477	100.0	

	CCHS Canadian Population		*Unweighted HRA Population (n=86,589)*		*Odds Ratio (95% CI)*
	#	%	#	%	
Some post-secondary					
25-34 yrs	375242	40.2	2229	18.4	0.34 (0.32-0.35)
35-44 yrs	124396	13.3	2094	17.3	1.36 (1.30-1.43)
45-64 yrs	300766	32.2	6389	52.8	2.36 (2.27-2.44)
65-89 yrs	132521	14.2	1378	11.4	0.78 (0.73-0.82)
Subtotal	932925	100.0	12989	100.0	
University/College graduate					
25-34 yrs	2208710	28.8	10680	22.1	0.07 (0.07-0.07)
35-44 yrs	1743154	22.8	10250	21.2	0.91 (0.89-0.93)
45-64 yrs	2812374	36.7	23376	48.3	1.61 (1.58-1.64)
65-89 yrs	892517	11.7	4098	8.5	0.70 (0.68-0.72)
Subtotal	7656755	100.0	78922	100.0	
Females – All Ages					
< High school	1847094	14.7	3181	4.0	0.24 (.023-0.25)
High school	2114807	16.8	15247	19.3	1.18 (1.16-1.20)
Some post-secondary	932925	7.4	12909	15.3	2.25 (2.21-2.30)
Univer/Coll graduate	7656755	61.0	48404	61.3	1.01 (0.99-1.03)*
Subtotal	122551581	100.0	78922	100.0	
Males & Females – All Ages					
<High school	3586810	14.6	5151	4.4	0.27 (0.26-0.38)
High school	4034829	16.4	21571	18.6	1.17 (1.15-1.19)
Some post-secondary	1828828	7.4	18103	15.6	2.31 (2.27-2.35)
Univer/Coll graduate	15170337	61.6	70990	61.3	0.88 (0.98-0-1.00)*
Total	24620904	100.0	115814	100.0	

*p>.001

Table 3: Highest level of education by age group and gender, CCHS (weighted) and HRA (weighted) populations

	Canadian Population		***Weighted HRA Population (n=86,589)***		***Odds Ratio (95% CI)***
	#	**%**	**#**	**%**	
Males					
< High School					
20-34 yrs	29079	16.7	274	14.5	0.85 (0.75-0.96)
35-44 yrs	194540	11.2	209	11.1	0.99 (0.86-1.14)*
45-64 yrs	633937	36.4	718	38.0	1.07 (0.97-1.17)*
65-89 yrs	621160	35.7	689	36.5	1.03 (0.94-1.13)*
Subtotal	1478716	100.0	1890	100.0	
High School					
20-34 yrs	569332	29.7	1362	22.1	0.67 (0.63-0.71)
35-44 yrs	361876	18.8	774	12.6	0.62 (0.57-0.67)
45-64 yrs	752165	39.2	2655	43.1	1.18 (1.12-1.24)
65-89 yrs	236749	12.3	1372	22.1	2.04 (1.92-2.16)
Subtotal	1920122	100.0	6163	100.0	
Some post-secondary					
25-34 yrs	401792	44.8	1645	15.7	0.46 (0.43-0.49)
35-44 yrs	122795	13.7	999	16.5	1.24 (1.16-1.33)
45-64 yrs	263371	29.4	2381	39.3	1.56 (1.48-1.64)
65-89 yrs	107945	12.0	1027	17.0	1.49 (1.39-1.60)
Subtotal	895903	100.0	6052	100.0	
University/College graduate					
25-34 yrs	2004798	26.7	7187	30.5	1.22 (1.18-1.25)
35-44 yrs	1659949	22.1	4809	20.4	0.91 (0.88-0.94)
45-64 yrs	2881497	38.4	8429	35.8	0.90 (0.88-0.93)
65-89 yrs	967338	12.9	3113	13.2	1.04 (1.00-1.08)
Subtotal	7513582	100.0	23438	100.0	
Males – All Ages					
< High School	1739716	32.8	1890	5.0	0.11 (0.10-0.11)
High School	1920211	36.2	6163	16.4	0.35 (0.34-0.36)
Some post-secondary	895903	16.9	6052	16.1	0.94 (0.92-0.97)
Univ/Coll graduate	751382	34.2	23538	62.5	10.11 (9.91-10.33)
Subtotal	5307212	100.0	37643	100.0	
Females					
< High School					
25-34 yrs	203649	11.0	972	24.5	2.62 (2.44-2.82)
35-44 yrs	135085	7.3	328	8.3	1.14 (1.02-1.28)*
45-64 yrs	611565	33.1	1204	30.4	0.88 (0.82-0.94)
65-89 yrs	896795	48.6	1463	36.9	0.62 (0.58-0.66)
Subtotal	1847094	100.0	3967	100.0	
High School					
25-34 yrs	469038	22.2	5691	31.2	1.59 (1.54-1.64)
35-44 yrs	332535	15.7	1726	9.5	0.56 (0.53-0.59)
45-64 yrs	893525	42.3	6457	35.4	0.75 (0.73-0.77)
54-89 yrs	419709	19.8	4364	23.9	1.27 (1.23-1.31)
Subtotal	2114807	100.0	18238	100.0	

	Canadian Population		*Weighted HRA Population (n=86,589)*		*Odds Ratio (95% CI)*
	#	%	#	%	
Some post-secondary					
25-34 yrs	375242	40.2	6343	41.2	1.07 (1.03-1.10)
35-44 yrs	124396	13.3	2003	13.0	0.97 (0.93-1.01)*
45-64 yrs	300766	32.2	6648	30.2	1.60 (1.55-1.65)
65-89 yrs	132521	14.2	2395	15.6	1.11 (1.07-1.16)
Subtotal	932925	100.0	15389	100.0	
University/College graduate					
25-34 yrs	2208710	28.8	30458	47.2	2.20 (2.17-2.24)
35-44 yrs	1743154	22.8	9814	15.2	0.61 (0.60-0.62)
45-64 yrs	2812374	36.7	17083	26.5	0.38 (0.37-0.39)
65-89 yrs	892517	11.7	7147	11.1	0.94 (0.92-0.97)
Subtotal	7656755	100.0	64502	100.0	
Females – All Ages					
< High school	1847094	14.7	3967	3.9	0.23 (.023-0.24)
High school	2114807	16.8	18238	17.9	1.07 (1.06-1.09)
Some post-secondary	932925	7.4	15389	15.1	2.21 (2.17-2.25)
Univer/Coll graduate	7656755	61.0	64502	63.2	1.10 (1.08-1.11)
Total	122551581	100.0	102096	100.0	
Males & Females – All Ages					
<High school	3586810	14.6	5857	4.2	0.26 (0.25-0.26)
High school	4034829	16.4	24401	17.5	1.08 (1.06-1.09)
Some post-secondary	1828828	7.4	21441	15.3	2.26 (2.23-2.29)
Univer/Coll graduate	15170337	61.6	88040	63.0	1.06 (1.05-1.07)
Total	24620904	100.0	139739	100.0	

* *p*>.001

Table 4: Select medical diagnoses by age, CCHS vs. HRA

	Arthritis						Diabetes					
	CCHS Population		HRA Population				CCHS Population		HRA Population			
	n	%	n	%	Odds Ratio (95% CI)	*p*	n	%	n	%	Odds Ratio (95% CI)	*p*
Males												
20-34	59509	1.8	157	2.5	1.57 (1.34-1.84)	<.001	39162	1.2	127	2.1	1.93 (1.62-2.30)	<.001
35-44	171576	7.2	404	5.9	0.80 (0.72-0.88)	<.001	75287	3.1	294	4.3	1.36 (1.21-1.53)	<.001
45-64	787304	16.7	2729	14.4	0.87 (0.84-0.91)	<.001	502245	10.6	2094	11.1	1.08 (1.03-1.13)	<.001
65+	676059	33.1	1613	25.6	0.75 (0.71-0.79)	<.001	433885	21.2	1127	17.9	0.86 (0.81-0.92)	<.001
Subtotal	1694448	13.1	4903	12.8	0.98 (0.96-1.01)	0.3	1050579	8.1	3642	9.5	1.18 (1.14-1.22)	<.001
Females												
20-34	99567	3	611	3.9	1.39 (1.28-1.51)	<.001	38635	1.2	297	1.9	1.74 (1.55-1.95)	<.001
35-44	197333	8.3	1245	8.3	0.99 (0.93-1.04)	0.62	58339	2.5	524	3.5	1.42 (1.30-1.55)	<.001
45-64	1169422	24.4	9896	23.9	1.01 (0.99-1.04)	0.31	312269	6.5	2834	6.8	1.09 (1.05-1.13)	<.001
65+	1263305	51.0	3791	41.8	0.75 (0.72-0.79)	<.001	368812	14.9	926	10.2	0.68 (0.64-0.73)	<.001
Subtotal	2729627	20.4	15543	19.2	0.94 (0.92-0.96)	<.001	778055	5.8	4581	5.6	0.97 (0.94-1.00)	0.06
Both sexes												
20-34	159076	2.4	768	3.5	1.60 (1.49-1.72)	<.001	77896	1.2	424	12.1	1.80 (1.63-1.98)	<.001
35-44	368909	7.7	1649	7.6	0.96 (0.91-1.01)	0.11	133626	2.8	818	10.7	1.33 (1.24-1.43)	<.001
45-64	1956726	20.6	12625	20.9	1.06 (1.04-1.08)	<.001	814516	8.6	4928	8.9	0.98 (0.95-1.01)	0.24
65+	1939364	43.0	5404	35.1	0.78 (0.76-0.81)	<.001	802696	17.7	2053	7.7	0.76 (0.72-0.79)	<.001
All	4424075	16.8	20446	17.2	1.02 (1.01-1.04)	0.004	1828734	6.9	8223	6.9	0.99 (0.97-1.02)	0.65

Table 4 (continued)

	Asthma						High Blood Pressure					
	CCHS Population		HRA Population				CCHS Population		HRA Population			
	n	%	n	%	Odds Ratio (95% CI)	*p*	n	%	n	%	Odds Ratio (95% CI)	*p*
Males												
20-34	286635	8.5	563	9.1	1.17 (1.08-1.28)	<.001	94854	2.8	534	8.7	3.55 (3.25-3.88)	<.001
35-44	150599	6.3	570	8.3	1.33 (1.22-1.45)	<.001	237921	10.0	1325	19.4	2.14 (2.01-2.27)	<.001
45-64	257280	5.5	1217	6.4	1.23 (1.16-1.31)	<.001	1179760	25.1	7185	37.9	1.91 (1.86-1.97)	<.001
65+	127687	6.2	383	6.1	0.99 (0.90-1.10)	0.89	892047	43.7	3487	55.3	1.74 (1.66-1.83)	<.001
Subtotal	822201	6.3	2733	7.1	1.13 (1.09-1.18)	<.001	2404582	18.5	12531	32.7	1.77 (1.74-1.81)	<.001
Females												
20-34	349622	10.5	2097	13.3	1.40 (1.34-1.46)	<.001	62098	1.9	798	5.1	2.97 (2.78-3.21)	<.001
35-44	216537	9.1	1766	11.8	1.31 (1.25-1.38)	<.001	145489	6.1	1833	12.3	2.11 (2.00-2.21)	<.001
45-64	498117	10.4	4139	10.0	0.99 (0.96-1.02)	0.53	1032601	21.5	11828	28.6	1.51 (1.48-1.55)	<.001
65+	189773	7.6	807	8.9	1.23 (1.15-1.33)	<.001	1265797	51.1	4480	49.4	1.02 (0.98-1.06)	0.35
Subtotal	1254049	9.4	8809	10.9	1.16 (1.13-1.18)	<.001	2505985	18.7	18939	23.4	1.25 (1.23-1.27)	<.001
Both sexes												
20-34	636257	9.5	2660	12.1	1.42 (1.36-1.48)	<.001	156952	2.3	1332	6.1	2.72 (2.57-2.87)	<.001
35-44	367135	7.7	2336	10.7	1.42 (1.36-1.48)	<.001	383409	8.0	3158	14.5	1.91 (1.84-1.98)	<.001
45-64	755397	7.9	5347	8.9	1.17 (1.13-1.20)	<.001	2212361	23.3	19013	31.5	1.58 (1.55-1.61)	<.001
65+	317460	7.0	1190	7.7	1.17 (1.10 - 1.24)	<.001	2157844	47.7	7967	51.8	1.29 (1.25-1.33)	<.001
All	2076249	7.9	11533	9.7	1.23 (1.21-1.25)	<.001	4910566	18.6	31470	26.4	1.42 (1.40-1.44)	<.001

Table 4 (continued)

	Smoking						Overweight and obese					
	CCHS Population		HRA Population				CCHS Population		HRA Population			
	n	%	n	%	Odds Ratio (95% CI)	*p*	n	%	n	%	Odds Ratio (95% CI)	*p*
Males												
20-34	1076445	32.0	1081	17.5	0.51 (0.47-0.54)	<.001	1608859	48.3	3447	55.9	1.60 (1.52-1.68)	<.001
35-44	657584	27.5	1022	14.9	0.45 (0.42-0.48)	<.001	1510783	64.3	4661	68.1	1.19 (1.13-1.26)	<.001
45-64	1237852	26.3	2478	13.1	0.44 (0.42-0.46)	<.001	3258305	70.3	13094	69.1	1.11 (1.08-1.15)	<.001
65+	240727	11.9	350	5.6	0.47 (0.42-0.52)	<.001	1168944	54.4	3973	63.0	1.43 (1.36-1.50)	<.001
Subtotal	3212608	27.4	4931	12.9	0.52 (0.51-0.54)	<.001	7546891	58.0	25175	65.7	1.13 (1.12-1.15)	<.001
Females												
20-34	733891	22.1	2380	15.1	0.68 (0.65-0.71)	<.001	964751	32.1	7946	50.5	2.72 (2.63-2.80)	<.001
35-44	457719	19.3	2170	14.6	0.70 (0.67-0.73)	<.001	894204	40.3	9080	60.9	2.50 (2.42-2.59)	<.001
45-64	946609	19.8	5010	12.1	0.58 (0.56-0.60)	<.001	2330192	51.1	26360	63.7	1.96 (1.92-3.00)	<.001
65+	240921	9.8	488	5.4	0.55 (0.51-0.61)	<.001	1175419	51.8	5527	60.9	1.88 (1.80-1.96)	<.001
Subtotal	2379140	17.8	10048	12.4	0.70 (0.68-0.71)	<.001	5364566	40.1	48913	60.3	1.51 (1.49-1.52)	<.001
Both sexes												
20-34	1810337	27.1	3461	15.8	0.56 (0.54-0.58)	<.001	2573610	40.6	11393	52.0	1.94 (1.88-1.99)	<.001
35-44	1115304	23.4	3192	14.7	0.55 (0.53-0.57)	<.001	2404987	52.7	13741	63.2	1.53 (1.58-1.67)	<.001
45-64	2184460	23.0	7488	12.4	0.49 (0.48-0.51)	<.001	5588496	60.8	39454	65.4	1.43 (1.40-1.45)	<.001
65+	481647	10.7	838	5.5	0.51 (0.48-0.55)	<.001	2344363	56.7	9500	61.8	1.65 (1.60-1.71)	<.001
All	5591748	21.2	14979	12.5	0.59 (0.58-0.60)	<.001	12911456	48.9	74088	62.1	1.27 (1.26-1.28)	<.001

Table 4 (continued)

	Mood Disorder						COPD					
	CCHS Population		HRA Population				CCHS Population		HRA Population			
	n	%	n	%	Odds Ratio (95% CI)	*p*	n	%	n	%	Odds Ratio (95% CI)	*p*
Males												
20-34	161783	4.8	640	10.4	2.49 (2.29-2.70)	<.001			61	1.0		
35-44	136330	5.7	855	12.5	2.32 (2.16-2.50)	<.001	41602	1.7	75	1.1	0.62 (0.49-0.77)	<.001
45-64	284099	6.0	2305	12.2	2.24 (2.14-2.34)	<.001	168982	3.6	392	2.1	0.59 (0.54-0.65)	<.001
65+	82911	4.1	516	8.2	2.22 (2.03-2.43)	<.001	144446	7.1	301	4.8	0.70 (0.62-0.78)	<.001
Subtotal	665123	5.1	4316	11.3	2.21 (2.14-2.28)	<.001	355030	2.7	829	2.2	0.79 (0.74-0.85)	<.001
Females												
20-34	250842	7.5	3096	19.7	3.21 (2.08-3.33)	<.001			115	0.7		
35-44	226104	9.5	3358	22.5	2.71 (2.60-2.81)	<.001	43793	1.8	130	0.9	0.46 (0.39-0.55)	<.001
45-64	183744	10.1	8248	19.9	6.44 (6.28-6.60)	<.001	220200	4.6	825	2.0	0.44 (0.41-0.47)	<.001
65+	157253	6.3	1236	13.6	2.44 (2.30-2.59)	<.001	186117	7.5	465	5.1	0.70 (0.64-0.77)	<.001
Subtotal	817943	6.1	15938	19.7	3.22 (3.16-3.27)	<.001	450110	3.4	1535	3.8	0.56 (0.54-0.59)	<.001
Both sexes												
20-34	412625	6.1	3736	17.1	3.36 (3.25-3.49)	<.001			176	0.8		
35-44	362435	7.6	4213	19.4	2.87 (2.77-2.96)	<.001	85394	1.8	205	0.9	0.51 (0.45-0.59)	<.001
45-64	767843	8.1	10553	17.5	2.50 (2.44-2.55)	<.001	389182	4.1	1217	2.0	0.50 (0.47-0.53)	<.001
65+	240153	5.3	1752	11.4	2.41 (2.29-2.53)	<.001	330563	7.3	766	5.0	0.70 (0.65-0.75)	<.001
All	1783056	6.8	20254	17.0	2.51 (2.48-2.55)	<.001	805139	3.1	2364	2.0	0.65 (0.62-0.68)	<.001

Table 5: Comparison of Non-Air Miles and Air Miles participants

	Non-Air Miles (n=48,056)	**Air Miles (n=72,454)**	**Both (n=120,510)**	**Effect size**
	Mean (sd)	**Mean (sd)**	**Mean (sd)**	**Cohen's *d***
Mean age in yrs	48.8 (14.1)	48.4 (14.2)	48.6 (14.1)	.028
Number of vascular diseases	0.7 (1.0)	0.5 (0.9)	0.6 (1.0)	.210
Number of non-modifiable risk factors	2.2 (1.5)	2.0 (1.4)	2.1 (1.5)	.138
Number of modifiable risk factors	2.7 (1.4)	2.5 (1.4)	2.6 (1.4)	.143
Total number of health concerns	5.6 (2.5)	5.0 (2.4)	5.3 (2.5)	.245
Total lifestyle healthiness score	29.3 (3.7)	28.7 (4.0)	28.9 (3.9)	.156
Demographics	**%**	**%**	**%**	**Cramer's V**
Male gender	31.9	32.1	32.0	.003 (p=.354)
Age group				
18-34	18.0	19.8	19.1	.eta=
35-44	17.2	18.6	18.1	.043
45-54	27.9	24.6	25.9	
55-64	24.4	24.0	24.1	
65-74	9.9	10.6	10.3	
75-90	2.6	2.3	2.4	
Education:				
Less than high school	4.9	4.1	4.4	.026
High school	19.7	18.6	18.7	
Some post-secondary	14.5	15.9	15.4	
College/university graduate	60.4	59.8	60.0	
Rather not say	1.5	1.6	1.6	
Employment:				
Full/part-time	59.2	56.4	57.5	.073
Self-employed	9.6	9.1	9.3	
Full-time student	5.1	3.3	4.0	
Stay-at-home parent	3.0	4.1	3.7	
Retired	16.8	19.9	18.7	
Other	6.4	7.1	6.8	
Type of work:				
Mgmt, health sciences, educ	58.1	63.0	65.0	.063
Sales or service	9.9	13.1	11.8	
Trades	9.3	8.8	9.0	
Rather not say	12.7	15.1	14.1	
Marital Status				
Married	58.9	57.9	58.3	.031
Common law	13.1	12.1	12.5	
Widowed, sep., div.	12.0	13.9	13.1	
Single/never married	14.7	14.7	14.7	
Rather not say	1.2	1.5	1.4	
High-risk ethnicity	6.8	4.6	5.5	.046
Age ≥55 years	36.9	36.9	36.9	.000
Fam Hx stroke	17.3	14.9	15.8	.031
Fam Hx heart dis.	41.8	34.9	37.6	.070
Fam Hx dyslipidemia	45.7	44.2	44.8	.015
Fam Hx diabetes	44.2	45.5	45.0	.013
Fam Hx hypertension	61.4	56.1	59.2	.053
Physically inactive	45.8	42.3	43.7	.035

	Non-Air Miles (n=48,056)	Air Miles (n=72,454)	Both (n=120,510)	Effect size
Smoking	11.1	13.5	12.5	.036
Overweight	69.1	61.8	64.8	.075
Excess alcohol	25.3	23.3	24.1	.023
Frequent fatty foods	15.2	11.7	13.1	.051
Frequent fast foods	3.4	2.5	2.9	.024
Low fruit/vegetable	42.6	41.9	42.1	.007 (p=.018)
Low fish consumption	53.3	53.1	53.2	.001 (p=.617)
≥1 bad diet behaviour	70.3	68.5	69.2	.020
Frequent stress	24.3	16.7	19.8	.093
High salt consumption	28.4	26.9	27.5	.016
Contemplation/Precontemplation				
Inactivity	45.8	42.3	43.7	.035
Smoking	7.0	11.3	9.6	.071
Alcohol	19.3	17.8	18.4	.019
Weight	47.8	57.1	53.4	.091
Diet	39.7	65.5	55.2	.254
Stress	15.9	16.2	16.1	.004 (p=.127)
Salt consumption	17.3	18.0	17.7	.009 (p=.002)
Diabetes	6.2	7.2	6.8	.019
Heart disease	5.4	3.8	4.4	.039
Hypertension	33.6	21.2	26.1	.138
Dyslipidemia	23.4	19.1	20.8	.052
Stroke	2.7	1.7	2.1	.033
Mood disorder	18.0	16.2	16.9	.023
Hypertension control				
Most of the time	49.9	72.8	61.0	.254
Some of the time	29.3	20.0	24.7	
Seldom/rarely	13.4	4.8	9.2	
Never	3.5	1.9	2.7	
Don't know	4.0	0.6	2.4	
Dyslipidemia control				
Most of the time	41.7	54.7	48.8	.186
Some of the time	28.6	24.6	25.5	
Seldom/rarely	14.4	9.7	11.8	
Never	7.4	8.2	7.9	
Don't know	9.9	2.8	6.0	
Diabetes control				
Most of the time	56.1	64.2	61.3	.147
Some of the time	27.8	24.3	25.6	
Seldom/rarely	8.2	6.1	6.9	
Never	2.9	4.4	3.8	
Don't know	5.0	1.0	2.4	
Poor medication adherence	44.4	41.0	42.2	.034
Entry portal				
H&S HRA	29.3	0	11.7	.991
Mobile	6.0	0	2.4	
BPAP	33.4	0	13.3	
eSupport	1.1	100	60.6	
HWAP	30.2	0	12.0	

	Non-Air Miles (n=48,056)	Air Miles (n=72,454)	Both (n=120,510)	Effect size
Enrolled eSupport	2.8	49.0	30.6	.492
Interacted eSupport	0.5	6.8	4.3	.150

Except where indicated, comparisons were statistically significant (p<.001) by independent t-test (continuous variables) or Chi squares (categorical variables). For Cohen's d, 0.2 = small effect, 0.5 medium effect, and 0.8 large effect. For Cramer's V for 1 degrees of freedom (two categories) 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect.

Appendix 5: Tables for Chapter 7

Table 1: Latent class analyses using number of vascular diseases and modifiable and non-modifiable risk factors

Number of Clusters	BIC(LL)	Number of Parameters	L^2	Degrees of Freedom	p-value	Classification Error Rate
2	1083489.5244	7	19687.920	568	2.1e-3718	0.2086
3	1080177.7446	11	16329.395	564	5.2e-3015	0.2504
4	1080207.4026	15	16312.308	560	2.3e-3014	0.2872
5	1080253.9777	19	16312.137	556	3.0e-3017	0.4953

Table 2: K-means Solution 1: Four-group k-mean cluster solution based on number of vascular diseases and non-modifiable and modifiable risk factors

	Overall	**Group 4 n=34,527 (29.0%)**	**Group 2 n=21,620 (18.2%)**	**Group 3 n=29,468 (24.8%)**	**Group 1 n=33,326 (28.0%)**	**Effect size**
Continuous variables/counts	***x* (sd)**	***x* (sd)**	***x* (sd)**	***x* (sd)**	***x* (sd)**	**ϖ**
Clustering variables						
Number vascular diseases	0.6 (1.0)	0.4 (0.7)	1.3 (1.2)	0.4 (0.7)	0.6 (0.9)	.339
Number modifiable risk factors	2.6 (1.4)	1.4 (0.9)	3.9 (1.0)	1.3 (0.9)	1.8 (0.8)	.825
Number non-modifiable risk factors	2.1 (1.5)	1.1 (0.8)	3.6 (0.9)	0.8 (0.7)	3.2 (0.8)	.747
Variables not used for clustering but related						
Number health concerns	5.3 (2.5)	5.1 (1.3)	8.8 (1.5)	2.4 (1.2)	5.6 (1.3)	.843
Healthiness score	28.9 (3.9)	26.2 (3.1)	25.6 (3.4)	32.2 (1.9)	30.9 (2.3)	.722
Distance between cases & cluster centroid	1.3 (0.6)	1.4 (0.6)	1.2 (0.5)	1.7 (0.7)	1.2 (0.5)	.301
Variables not related to clustering						
Age in years	48.5 (14.1)	46.4 (13.8)	47.8 (13.1)	51.3 (14.8)	48.8 (14.1)	.128
Median age	50	47	49	53	50	
Categorical variables:	**%**	**%**	**%**	**%**	**%**	**Cramer's V**
Risk factors and vascular diseases related to clustering variables						
Physical inactivity	43.7	65.8	74.6	14.4	26.7	.497
Smoking	12.6	20.9	23.4	3.3	5.1	.266
Excess alcohol	24.1	36.5	34.9	12.2	14.7	.263
Fatty foods	13.1	18.4	21.1	5.9	8.7	.183
Fast foods	2.9	4.5	5.9	0.7	1.1	.127
Low fruit/vegetables	42.1	58.6	62.0	21.7	30.1	.348
Frequent stress	19.8	28.7	41.3	4.4	10.1	.348
Low fish consumption	53.2	68.8	70.5	33.2	43.5	.317
≥1 bad diet behaviour	69.6	90.0	91.0	44.2	56.9	.440
Overweight/obese	61.8	77.1	83.7	37.7	53.2	.368
Excess salt	27.5	45.5	44.5	9.9	13.3	.375
Unwilling to change:						
Inactive	52.8	57.0	56.3	40.7	44.9	.123
Smokers	37.9	38.9	40.0	28.8	31.9	.070
Excess alcohol	45.2	46.8	45.3	42.7	42.4	.036
Overweight	58.6	47.6	45.9	33.0	33.0	.139
Poor diet	24.9	30.9	30.8	13.1	17.0	.176
Stressed	32.3	35.2	34.6	16.8	23.5	.117
High salt	26.1	27.7	21.5	30.5	27.3	.070
Family history premature heart disease	37.4	18.2	68.2	13.9	58.3	.480
Family history	15.8	5.1	34.9	3.6	25.2	.348

	Overall	Group 4 n=34,527 (29.0%)	Group 2 n=21,620 (18.2%)	Group 3 n=29,468 (24.8%)	Group 1 n=33,326 (28.0%)	Effect size
premature stroke						
Family history dyslipidemia	44.6	21.9	80.4	15.4	70.9	.565
Family history hypertension	58.0	35.8	91.7	27.9	85.8	.570
Family history diabetes	44.9	25.1	76.4	18.7	67.9	.498
Higher risk ethnicity	5.5	3.2	10.8	2.1	7.4	.142
Diabetes	6.8	3.5	17.3	3.1	6.9	.205
Heart disease	4.4	2.2	9.7	3.0	4.6	.129
Hypertension	26.0	19.7	49.2	16.0	26.5	.264
Dyslipidemia	20.8	13.7	42.2	11.7	22.3	.269
Renal disease	1.3	0.7	3.1	0.9	1.0	.077
Stroke	2.1	1.1	4.9	1.5	1.8	.096
Proportion of those with diagnosis who have condition controlled "most of the time"						
Blood pressure	61.6	52.5	55.0	70.1	70.4	.107
Blood lipids	48.8	42.1	42.1	55.3	57.0	.085
Blood sugar	61.2	55.0	55.0	73.5	72.5	.124
Variables not related to clustering variables						
Prescribed medication	42.3	38.7	60.9	32.4	43.0	.193
Mood disorder	16.9	19.1	28.7	9.3	13.9	.176
Most/some of the time miss taking medication†	12.4	14.5	16.2	8.6	9.7	.095
Demographics						
Age Groups						
18-34	19.2	22.2	18.2	16.6	18.8	.087
35-44	18.1	20.7	20.0	14.3	17.5	
45-54	25.9	27.0	28.3	23.0	25.9	
55-64	24.1	20.8	24.1	27.1	24.9	
65-74	10.3	21.9	7.9	14.8	10.5	
75-90	2.4	1.5	1.4	4.3	2.3	
Age ≥65	9.3	9.3	9.3	19.1	12.8	.119
Entry portal/source:						
HRA landing page	11.6	11.5	11.1	12.2	11.5	.065
Mobile phone app	2.4	2.6	2.3	2.4	2.3	
BPAP	13.2	12.5	16.4	11.9	13.1	
eSupport	60.7	59.9	53.2	65.6	62.1	
HWAP	12.0	13.4	17.8	7.9	10.9	
Air Miles participant	60.3	59.6	52.6	65.2	61.6	.085
Enrollment for follow up:						
eSupport emails	30.7	30.5	29.4	30.9	31.7	.016
BPAP self-management	1.8	1.6	3.1	1.1	1.7	.051
HWAP	7.3	8.3	11.0	4.3	6.5	.088

	Overall	Group 4 n=34,527 (29.0%)	Group 2 n=21,620 (18.2%)	Group 3 n=29,468 (24.8%)	Group 1 n=33,326 (28.0%)	Effect size
Joined any	38.3	39.1	41.6	35.4	38.5	.043
Male gender	32.0	35.5	32.0	33.9	26.8	.074
Higher education	76.6	74.9	72.1	79.8	78.4	.067
Married	58.2	55.6	55.1	61.9	59.8	.057
Work full/part-time	58.6	62.1	60.3	53.0	58.7	.070
White collar occupation	65.0	63.9	61.7	66.1	67.4	.043

Missing = 1,038 † of those prescribed ≥1 medication * related to clustering variable(s)
For ϖ standard cut-offs are: 0.01 = small effect, 0.06 = medium effect and 0.14 =large effect.
For Cramer's V for 1 degrees of freedom, 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect. For Cramer's V with 3df; .06=small, .17 = moderate and .29 = large effect. All effect sizes significant (p<.001) unless stated otherwise.

Table 3: Two-step Solution 1: Four-group two-step cluster solution based on number of vascular diseases and non-modifiable and modifiable risk factors

	Overall	**Group 4 n=49,594 (41.7%)**	**Group 3 n=26,980 (22.7%)**	**Group 2 n=24,343 (20.5%)**	**Group 1 n=18,024 (15.2%)**	**Effect size**
Continuous variables/counts	***x* (sd)**	***x* (sd)**	***x* (sd)**	***x* (sd)**	***x* (sd)**	**ϖ**
Clustering variables						
Number vascular diseases	0.6 (1.0)	0.3 (0.4)	0.4 (0.5)	0.1 (0.3)	2.5 (0.8)	.340
Number modifiable risk factors	2.6 (1.4)	3.7 (0.9)	1.6 (0.7)	1.2 (0.7)	2.7 (1.3)	.843
Number non-modifiable risk factors	2.1 (1.5)	2.0 (1.4)	3.0 (1.1)	0.7 (0.7)	2.8 (1.4)	.547
Variables not used for clustering but related						
Number health concerns	5.3 (2.5)	6.0 (1.8)	5.1 (1.3)	2.1 (0.9)	8.0 (2.2)	.750
Healthiness score	28.9 (3.9)	26.0 (3.2)	31.3 (2.0)	32.3 (1.9)	28.7 (3.8)	.685
Variables not related to clustering						
Age in years	48.5 (14.1)	44.3 (13.2)	48.9 (13.6)	49.3 (14.6)	58.6 (11.1)	.340
Median age	50	45	50	51	59	
Categorical variables:	**%**	**%**	**%**	**%**	**%**	**Cramer's V**
Risk factors and vascular diseases related to clustering variables						
Physical inactivity	43.7	67.3	23.1	13.7	50.2	.460
Smoking	12.6	21.7	4.2	3.5	12.0	.248
Excess alcohol	24.1	35.3	13.9	11.9	25.0	.242
Fatty foods	13.1	20.0	7.8	6.0	11.5	.179
Fast foods	2.9	5.1	0.8	0.7	2.7	.119
Low fruit/vegetables	42.1	59.5	27.3	21.5	44.1	.329
Frequent stress	19.8	33.4	8.5	4.5	19.6	.312
Low fish consumption	53.2	70.0	41.2	33.6	51.5	.303
≥1 bad diet behaviour	69.6	90.5	53.9	44.4	69.3	.417
Overweight/obese	61.8	77.0	49.8	34.7	75.0	.360
Salt	27.5	10.7	11.3	20.3	47.2	.379
Unwilling to change:						
Inactive	52.8	57.0	43.2	40.9	51.1	.119
Smokers	37.8	39.6	30.5	28.3	36.6	.071
Excess alcohol	45.2	46.1	42.3	41.8	46.0	.034
Overweight	41.4	47.5	31.8	33.3	38.9	.137
Poor diet	24.9	31.5	15.0	13.2	22.9	.178
Stressed	32.3	35.8	20.1	17.1	28.4	.129
High salt	26.1	26.8	26.3	31.9	17.3	.077
Excess salt	27.5	47.2	11.3	10.7	20.3	.379
Family history premature heart disease	37.4	35.0	55.1	12.7	51.0	.311
Family history	15.8	14.6	25.2	3.4	21.7	.208

	Overall	Group 4 n=49,594 (41.7%)	Group 3 n=26,980 (22.7%)	Group 2 n=24,343 (20.5%)	Group 1 n=18,024 (15.2%)	Effect size
premature stroke						
Family history dyslipidemia	44.6	41.4	64.8	13.4	65.4	.382
Family history hypertension	58.0	55.0	81.6	23.8	77.4	.418
Family history diabetes	44.9	44.1	63.4	17.9	55.7	.314
Higher risk ethnicity	5.5	5.8	7.5	2.1	6.0	.080
Diabetes	6.8	1.5	2.3	0.6	36.7	.501
Heart disease	4.4	0.7	1.3	0.7	24.4	.410
Hypertension	26.0	15.4	23.2	5.7	87.1	.603
Dyslipidemia	20.8	9.2	15.3	3.7	83.9	.664
Renal disease	1.3	0.2	0.4	0.2	7.1	.218
Stroke	2.1	0.4	0.6	0.3	11.1	.269
Proportion of those with diagnosis who have condition controlled "most of the time						
Blood pressure	61.1	47.0	64.5	69.0	66.0	.107
Blood lipids	48.8	32.8	43.3	49.8	55.5	.115
Blood sugar	61.2	55.0	71.5	82.6	60.3	.060
Variables not related to clustering variables						
Mood disorder	16.9	21.2	12.9	8.7	22.3	.145
Prescribed medication	42.3	36.8	41.7	23.4	84.3	.380
Most/some of the time miss taking medication †	12.4	16.1	10.3	9.4	10.5	.087
Demographics						
Age Groups						
18-34	19.2	25.8	17.7	19.4	2.7	.200
35-44	18.1	23.4	17.3	16.1	7.4	
45-54	25.9	27.1	27.8	24.2	22.3	
55-64	24.1	17.8	25.6	25.4	37.7	
65-74	10.3	5.1	9.6	12.1	23.4	
75-90	2.4	0.9	2.0	2.9	6.5	
Age ≥65	12.7	5.9	11.6	15.0	29.9	.243
Entry portal/source:						
HRA landing page	11.6	11.4	11.9	11.5	11.9	.070
Mobile phone app	2.4	2.5	2.4	2.4	2.4	
BPAP	13.2	12.3	13.7	19.3	13.7	
eSupport	60.7	59.4	61.6	53.6	61.6	
HWAP	12.0	14.5	10.3	13.3	10.3	
Air Miles participant	60.3	59.0	61.2	67.1	53.1	.087
Enrollment:						
eSupport emails	30.7	30.8	32.0	28.4	31.0	.023
BPAP self-management	1.8	1.6	0.7	3.8	1.6	.070
HWAP	7.3	9.1	4.1	8.5	6.0	.077
Joined any	38.5	40.3	37.3	36.1	38.5	.035
Male gender	32.0	30.5	25.8	31.4	46.4	.137
Higher education	76.6	76.1	79.3	81.7	67.7	.100

	Overall	Group 4 n=49,594 (41.7%)	Group 3 n=26,980 (22.7%)	Group 2 n=24,343 (20.5%)	Group 1 n=18,024 (15.2%)	Effect size
Married	58.2	53.8	60.7	60.6	63.7	.080
Work full/part-time	58.6	65.2	60.0	56.0	41.7	.161
White collar occupation	65.0	65.3	68.7	67.7	55.3	.091

Missing = 1,038 † of those prescribed medication * related to clustering variable(s)
For ϖ standard cut-offs are: 0.01 = small effect, 0.06 = medium effect and 0.14 =large effect.
For Cramer's V for 1 degrees of freedom, 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect. For Cramer's V with 3df; .06=small, .17 = moderate and .29 = large effect . All effect sizes significant (p<.001) unless stated otherwise.

Table 4: Latent class analyses using healthiness score and number of health concerns

Number of Clusters	BIC(LL)	Number of Parameters	L^2	Degrees of Freedom	p-value	Classification Error Rate
2	1225563.3124	5	93407.336	734	7.5e-1935 6	0.2179
3	1225360.9821	8	93169.946	731	7.3e-1930 8	0.1932
4	1225348.8762	11	93122.781	728	7.3e-1930 1	0.3603
5	1225383.9336	14	93122.779	725	5.0e-1930 4	0.4332

Table 5: K-means Solution 2: Four-group solution based on number of health concerns and overall lifestyle healthiness score

	Overall	**Cluster 3 n=29,378 (24.7%) Younger & healthier**	**Cluster 2 n=16,715 (14.1%) Younger & less healthy**	**Cluster 1 n=46,797 (39.3%) Older & healthier**	**Cluster 4 n=26,051 (21.9%) Older & less healthy**	**Effect Size**
Continuous variables/counts	**Overall Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**ϖ**
Variables used for clustering						
Health concerns	5.3 (2.5)	4.8 (1.4)	7.9 (2.0)	3.4 (1.6)	7.5 (1.6)	.764
Overall healthiness score	**28.9 (3.9)**	27.0 (1.5)	22.1 (2.4)	32.4 (1.6)	29.3 (1.7)	.895
Variables not used in clustering but related						
Number of vascular disease	0.6 (1.0)	0.2 (0.5)	0.8 (1.1)	0.4 (0.7)	1.4 (1.2)	.470
Number of non-modifiable risk factors	2.1 (1.5)	1.4 (1.1)	2.5 (1.4)	1.6 (1.3)	3.3 (1.2)	.511
Number of modifiable risk factors	2.6 (1.4)	3.2 (0.7)	4.6 (0.8)	1.4 (0.8)	2.8 (0.9)	.803
Variable not used in clustering						
Age in yrs	45.5 (14.1)	44.8 (13.5)	45.2 (13.1)	49.5 (14.4)	53.3 (13.3)	.166
Median age	50	45	46	51	56	
Categorical variables	**Overall**	**%**	**%**	**%**	**%**	**Cramer's V**
Risk factors and vascular diseases related to clustering variables						
Physical inactivity	43.7	60.8	85.2	14.4	50.3	.518
Smoking	12.6	15.0	35.6	3.8	10.7	.312
Excess alcohol	24.1	29.3	45.4	12.6	25.2	.259
Fatty food	13.1	16.6	25.6	6.6	12.6	.191
Fast foods	2.9	3.6	8.5	0.7	2.2	.153
Low fruit/vegetable	42.1	53.5	71.4	22.9	33.9	.350
Low fish consumption	53.2	64.2	65.4	36.4	56.8	.293
Salt	27.5	37.0	60.9	11.6	24.0	.376
≥1 bad diet behaviour	69.6	83.7	95.8	47.9	75.8	.401
Overweight/obese	61.8	70.6	84.9	40.6	75.3	.363
Frequent stress	19.8	21.5	49.4	6.9	21.9	.347
Unwilling to change:						
Inactive	52.8	63.0	78.4	23.9	33.8	.415
Smokers	37.8	38.2	55.4	11.7	16.6	.350
Excess alcohol	45.2	54.6	63.0	24.9	30.3	.315
Overweight	41.4	54.8	75.4	19.0	24.5	.445
Poor diet	24.9	33.0	59.1	5.4	9.3	.467
Stressed	32.3	33.3	55.9	8.7	10.4	.422
High salt	26.1	32.8	37.2	11.7	8.6	.276
Family history hypertension	58.0	41.7	68.0	49.0	86.1	.341

	Overall	Cluster 3 n=29,378 (24.7%) Younger & healthier	Cluster 2 n=16,715 (14.1%) Younger & less healthy	Cluster 1 n=46,797 (39.3%) Older & healthier	Cluster 4 n=26,051 (21.9%) Older & less healthy	Effect Size
Family history of dyslipidemia	55.4	29.1	56.1	35.2	71.8	.333
Family history diabetes	44.9	32.3	55.0	36.1	86.1	.287
Family history premature heart disease	37.4	22.6	45.4	29.2	63.9	.322
Family history of premature stroke	15.8	7.5	21.6	10.8	30.3	.242
Higher risk ethnicity	5.5	3.7	7.4	4.0	8.9	.095
Diabetes	6.8	1.8	9.7	2.8	17.9	.253
Heart disease	4.4	1.0	4.9	2.3	11.9	.200
Hypertension	26.0	12.1	31.0	16.6	55.5	.379
Dyslipidemia	20.8	8.4	26.5	12.1	46.6	.364
Renal disease	1.3	0.3	1.8	0.5	3.4	.110
Stroke	2.1	0.5	2.6	1.0	5.4	.131
Proportion of those with diagnosis who have condition controlled "most of the time"						
Blood pressure	61.1	56.2	50.3	67.3	62.9	.078
Blood lipids	48.8	39.3	39.3	53.3	52.3	.082
Blood sugar	61.2	59.7	45.4	75.2	62.8	.121
Variables not related to the clustering variables						
Prescribed medication	42.3	32.6	50.3	23.1	33.5	.269
Most/some of time miss medication †	12.4	14.5	29.5	10.4	11.2	.102
Mood disorder	16.9	16.8	29.5	10.4	10.9	.174
Demographics						
Entry portal/source:						
HRA landing page	11.6	10.2	9.8	12.9	12.2	.095
Mobile phone app	2.4	2.0	1.6	2.9	2.6	
BPAP	13.2	8.7	10.2	13.7	19.3	
eSupport	60.7	68.4	67.1	60.1	49.1	
HWAP	12.0	10.7	11.3	10.4	16.8	
Air Miles participant	60.3	68.0	66.6	59.7	48.5	.144
Enrollment:						
eSupport emails	30.7	34.7	36.3	29.0	25.9	.082
BPAP self-mgmt.	1.8	1.1	1.7	1.3	3.4	.068
HWAP	7.3	7.0	7.7	5.5	10.5	.072
Joined any	38.5	41.3	44.1	35.0	37.8	.070

	Overall	Cluster 3 n=29,378 (24.7%) Younger & healthier	Cluster 2 n=16,715 (14.1%) Younger & less healthy	Cluster 1 n=46,797 (39.3%) Older & healthier	Cluster 4 n=26,051 (21.9%) Older & less healthy	Effect Size
Age groups:						
18-34	19.2	25.1	23.5	18.3	11.1	.144
35-44	18.1	22.8	23.2	16.3	12.8	
45-54	25.9	28.6	26.7	26.2	21.8	
55-64	24.1	16.1	20.1	24.3	35.4	
65-74	10.3	6.1	5.6	11.7	15.5	
75-90	2.4	1.3	0.8	3.2	3.3	
Age ≥65	9.3	7.4	6.4	14.9	18.8	.143
Male gender	32.0	32.7	34.0	30.1	33.6	.035
Higher education	76.6	77.8	72.2	79.8	72.2	.079
Married	58.2	55.4	51.3	60.9	61.1	.074
Work full/part-time	58.6	64.7	62.3	57.1	51.9	.095
White collar occupation	65.0	65.9	61.0	67.8	61.8	.059

Missing=1,569 † Of those prescribed ≥1 medication * related to clustering variable(s)

For ϖ standard cut-offs are: 0.01 = small effect, 0.06 = medium effect, and 0.14 =large effect. For Cramer's V for 1 degrees of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect. For Cramer's V with 3_{df}; -.06=small effect, 0.17 = moderate effect, and 0.29 = large effect. All effect sizes significant (p<.001) unless stated otherwise.

Table 6: Two-step Solution 2: Four-group two-step solution using healthiness scores and number of health concerns

	Overall	**Cluster 4 n=29709 (25.0%)**	**Cluster 3 n=20352 (17.1%)**	**Cluster 1 n=41272 (34.7%)**	**Cluster 2 n=27608 (23.2%)**	**Effect Size**
Continuous variables/counts	**Overall Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**ϖ**
Variables used for clustering						
Health concerns	5.3 (2.5)	5.4 (1.1)	8.9 (1.5)	2.8 (1.1)	6.2 (1.2)	.876
Overall healthiness score	28.9 (3.9)	26.1 (2.1)	24.2 (3.5)	31.8 (2.1)	31.1 (1.5)	.807
Variables not used in clustering but related						
Number of vascular disease	0.6 (1.0)	0.3 (0.6)	1.3 (1.3)	0.2 (0.5)	1.0 (1.1)	.466
Number of modifiable risk factors	2.6 (1.4)	3.4 (0.8)	4.3 (1.0)	1.5 (0.9)	2.1 (0.9)	.772
Number of non-modifiable risk factors	2.1 (1.5)	1.6 (1.1)	3.3 (1.2)	1.1 (1.0)	3.1 (1.1)	.669
Variable not used in clustering						
Distance between cases & cluster centroid	2.2 (1.2)	2.8 (1.6)	2.0 (1.0)	2.0 (0.9)	1.9 (0.8)	.257
Age in years	45.5 (14.1)	45.7 (13.8)	47.5 (13.2)	49.3 (14.6)	51.3 (13.9)	.144
Median age	50	46	49	51	53	
Categorical variables	**Overall**	**%**	**%**	**%**	**%**	**Cramer's V**
Risk factors and vascular diseases (related to clustering variables)						
Physical inactivity	43.7	66.6	80.5	18.6	29.4	.506
Smoking	12.6	17.7	29.5	4.8	6.1	.280
Excess alcohol	24.1	32.3	40.0	14.2	18.2	.236
Fatty food	13.1	17.9	22.6	7.2	9.6	.178
Fast foods	2.9	4.1	7.2	0.9	1.3	.141
Low fruit/vegetable	42.1	57.2	65.8	25.2	33.6	.332
Low fish consumption	53.2	67.1	71.9	37.8	47.5	.283
Salt	27.5	40.7	51.2	13.8	16.2	.343
≥1 bad diet behaviour	69.6	87.1	92.5	49.9	63.2	.383
Overweight/obese	61.8	74.7	85.4	40.7	62.2	.352
Frequent stress	19.8	25.4	46.0	6.5	14.0	.350
Unwilling to change:						
Inactive	52.8	65.8	64.8	37.5	21.9	.355
Smokers	37.8	42.4	46.0	22.9	12.4	.241
Excess alcohol	45.2	55.8	52.4	37.2	22.4	.251
Overweight	41.4	57.4	56.7	30.1	16.5	.355
Poor diet	24.9	37.5	40.7	11.6	5.0	.354
Stressed	32.3	38.7	42.3	15.7	7.2	.294
High salt	26.1	34.1	26.6	23.1	7.1	.197

	Overall	Cluster 4 n=29709 (25.0%)	Cluster 3 n=20352 (17.1%)	Cluster 1 n=41272 (34.7%)	Cluster 2 n=27608 (23.2%)	Effect Size
Family history hypertension	58.0	48.9	84.7	34.0	84.3	.455
Family history of dyslipidemia	44.6	34.0	73.9	21.8	68.7	.445
Family history diabetes	44.9	36.9	70.4	24.6	65.0	.384
Family history premature heart disease	37.4	26.6	62.5	18.3	59.3	.400
Family history of premature stroke	15.8	8.7	32.1	5.6	26.6	.302
Higher risk ethnicity	5.5	3.9	10.3	2.5	8.0	.132
Diabetes	6.8	2.6	18.5	1.3	11.1	.258
Heart disease	4.4	1.4	10.4	1.2	8.1	.188
Hypertension	26.0	16.5	48.9	10.0	43.6	.376
Dyslipidemia	20.8	11.7	42.5	6.9	35.2	.363
Renal disease	1.3	0.3	3.6	0.3	2.1	.115
Stroke	2.1	0.6	5.4	0.6	3.4	.134
Proportion of those with diagnosis who have condition controlled "most of the time"						
Blood pressure	61.1	58.2	54.7	68.2	65.1	.069
Blood lipids	48.8	42.8	43.3	52.5	55.0	.074
Blood sugar	61.2	59.4	52.2	75.1	70.0	.119
Variables not related to the clustering variables						
Prescribed medication	42.3	37.9	60.9	27.7	55.4	.270
Most/some of time miss medication †	12.4	13.7	17.3	9.6	9.7	.097
Mood disorder	16.9	18.6	30.7	9.9	15.6	.190
Demographics						
Entry portal/source:						
HRA landing page	11.6	10.0	10.7	12.3	13.1	.094
Mobile phone app	2.4	1.9	2.0	2.5	3.1	
BPAP	13.2	9.6	14.2	11.3	19.2	
eSupport	60.7	67.7	57.9	64.7	49.2	
HWAP	12.0	10.8	15.1	9.2	15.3	
Air Miles	60.3	67.3	57.4	64.3	48.8	.144
Enrollment:						
eSupport emails	30.7	34.9	31.9	30.7	25.4	.073
BPAP self-mgmt	1.8	1.3	2.8	1.0	2.8	.064
HWAP	7.3	7.2	10.1	5.0	8.8	.074
Joined any	38.5	41.9	42.8	35.8	35.5	.068

	Overall	Cluster 4 n=29709 (25.0%)	Cluster 3 n=20352 (17.1%)	Cluster 1 n=41272 (34.7%)	Cluster 2 n=27608 (23.2%)	Effect Size
Age groups:						
18-34	19.2	23.9	18.9	19.1	14.4	.155
35-44	18.1	21.6	20.5	16.4	15.1	
45-54	25.9	26.1	28.5	24.8	25.4	
55-64	24.1	19.9	23.1	24.9	28.3	
65-74	10.3	7.2	7.6	11.8	13.4	
75-90	2.4	1.3	1.4	3.1	3.3	
Age ≥65	9.3	8.5	9.0	14.8	16.8	.105
Male gender	32.0	32.9	33.8	31.1	31.2	.023
Higher education	76.6	76.3	71.1	80.1	75.6	.074
Married	58.2	55.3	53.5	60.5	61.0	.066
Work full/part-time	58.6	63.0	59.6	56.8	55.7	.058
White collar occupation	65.0	64.9	60.4	67.5	65.0	.051

Missing=1,569 † Of those prescribed ≥1 medication * related to clustering variable(s)

For ϖ standard cut-offs are: 0.01 = small effect, 0.06 = medium effect, and 0.14 =large effect. For Cramer's V for 1 degrees of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect. For Cramer's V with 3_{df}; -.06=small effect, 0.17 = moderate effect, and 0.29 = large effect. All effect sizes significant (p<.001) unless stated otherwise.

Table 7: K-means Solution 3: Four-group solution based on age, lifestyle healthiness score, number of vascular diseases and number of non-modifiable risk factors

	Overall	Cluster 1 n=24,643 (20.7%)	Cluster 4 n=31,563 (26.5%)	Cluster 3 n=40,532 (34.0%)	Cluster 2 n=22,346 (18.8%)	Effect Size
Continuous variables/counts	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**ϖ**
Variables used for clustering						
Age in years	48.5 (14.1)	28.0 (4.8)	42.7 (3.8)	54.9 (3.6)	68.0 (5.3)	.953
Median age	50	28	43	55	67	
Lifestyle healthiness score	28.9 (3.9)	28.2 (4.1)	28.2 (4.1)	29.3 (3.7)	30.1 (3.3)	.193
Number of vascular diseases	0.6 (1.0)	0.2 (0.6)	0.4 (0.8)	0.7 (1.0)	1.2 (1.1)	.362
Number non-modifiable risk factors	2.1 (1.5)	2.1 (1.5)	2.2 (1.5)	2.1 (1.4)	1.9 (1.4)	.072
Distance of cases from cluster centre	5.5 (2.4)	6.1 (2.6)	6.2 (2.6)	5.0 (2.0)	5.7 (2.8)	.058
Variables not used in clustering but related						
Number of modifiable risk factors	2.6 (1.4)	2.9 (1.4)	2.9 (1.4)	2.5 (1.3)	2.1 (1.2)	.207
Total number health concerns	5.3 (2.5)	5.1 (2.3)	5.4 (2.5)	5.3 (2.6)	5.1 (2.5)	.166
Categorical variables	**Overall**	**%**	**%**	**%**	**%**	**Cramer's V**
Risk factors, vascular diseases and age groups related to clustering variables						
Physical inactivity	43.7	52.7	50.0	58.4	65.5	.112
Smoking	12.6	15.6	15.2	12.1	6.1	.103
Excess alcohol	24.1	25.1	22.2	24.6	23.2	.019
Fatty food	13.1	20.2	15.2	10.1	7.6	.133
Fast foods	2.9	6.1	3.6	1.6	0.6	.116
Low fruit/vegetable	42.1	49.9	47.1	37.8	34.4	.123
Frequent stress	19.8	25.3	25.2	18.1	8.8	.154
Low fish consumption	53.2	58.7	59.9	50.7	42.5	.130
≥1 bad diet behaviour	69.6	76.9	75.6	66.2	59.2	.146
Overweight/obese	61.8	52.1	64.3	65.1	63.3	.103
Salt	27.5	42.9	32.4	20.9	15.5	.221
Unwilling to change:						
Inactive	52.8	53.2	55.6	51.0	50.4	.041
Smokers	37.8	38.2	41.4	36.2	26.9	.067
Excess alcohol	45.2	37.1	46.2	48.6	46.9	.088
Overweight	41.4	43.4	44.4	39.1	39.8	.048
Poor diet	24.9	29.4	27.9	21.6	19.7	.090
Stressed	32.3	37.7	34.7	27.3	23.8	.104
High salt	26.1	31.5	26.3	21.3	20.2	.099
Family history premature heart	37.4	32.1	38.8	40.0	36.7	.062

	Overall	Cluster 1 n=24,643 (20.7%)	Cluster 4 n=31,563 (26.5%)	Cluster 3 n=40,532 (34.0%)	Cluster 2 n=22,346 (18.8%)	Effect Size
disease						
Family history of premature stroke	15.8	15.7	16.9	15.9	14.0	.027
Family history dyslipidemia	44.6	47.2	46.6	45.2	38.1	.065
Family history hypertension	58.0	55.8	59.0	59.9	55.7	.038
Family history diabetes	44.0	48.2	47.0	44.8	38.3	.069
Higher risk ethnicity	5.5	8.5	7.3	3.7	2.7	.101
Diabetes	6.8	2.0	4.5	8.2	12.9	.148
Heart disease	4.4	1.0	1.9	4.6	11.4	.177
Hypertension	26.0	6.2	17.6	32.1	48.8	.330
Dyslipidemia	20.8	5.0	13.8	26.2	38.2	.282
Renal disease	1.3	0.9	1.1	1.3	2.0	.031
Stroke	2.1	0.8	1.4	2.1	4.4	.085
Proportion of those with diagnosis who have condition controlled "most of the time"						
Blood pressure	61.1	35.2	46.2	61.5	72.2	.144
Blood lipids	48.8	24.9	32.3	48.0	62.6	.155
Blood sugar	61.2	47.7	53.8	60.0	68.4	.109
Variables not related to the clustering variables						
Prescribed medication	42.3	22.4	33.8	48.3	65.5	.297
Most/some of time miss medication†	12.4	21.8	16.7	11.2	7.1	.144
Mood disorder	16.9	17.1	19.2	17.6	12.4	.062
Demographics						
Age groups:*						
18-34	19.2	92.6	0	0	0	.835
35-44	18.1	7.4	62.4	0	0	
45-54	25.9	0	37.6	36.9	0	
55-64	24.1	0	0	53.1	32.3	
65-74	10.3	0	0	0	54.9	
75-90	2.4	0	0	0	12.8	
Age ≥65 *	12.7	0	0	0	67.7	.794
Entry portal/source:						
HRA landing page	11.6	11.2	11.2	12.3	11.2	.054
Mobile phone app	2.4	3.2	2.9	2.0	1.6	
BPAP	13.2	9.0	12.5	14.9	15.8	
eSupport	60.7	63.2	61.2	58.1	61.9	
HWAP	12.0	13.4	12.0	12.7	9.3	
Air Miles participant	60.3	62.6	60.8	57.8	61.6	.039
Enrollment:						
eSupport emails	30.7	32.8	32.2	29.8	28.1	.038
BPAP self-mgmt	1.8	0.7	1.7	2.2	2.3	.045
HWAP	7.3	6.8	7.4	8.3	6.0	.032
Joined any	38.4	39.7	40.3	38.3	34.6	.041

	Overall	Cluster 1 n=24,643 (20.7%)	Cluster 4 n=31,563 (26.5%)	Cluster 3 n=40,532 (34.0%)	Cluster 2 n=22,346 (18.8%)	Effect Size
Male gender	68.0	71.6	68.4	69.6	60.4	.081
Higher education	76.6	82.2	82.7	74.0	66.6	.144
Married	58.3	33.1	62.4	65.7	66.5	.262
Work full/part-time	58.5	66.7	75.9	63.3	17.0	.420
White collar occupation	65.0	66.9	71.0	67.0	50.8	.147

Missing=1,424 † Of those prescribed ≥1 medication * related to clustering variable(s)

For ϖ standard cut-offs are: 0.01 = small effect, 0.06 = medium effect, and 0.14 =large effect. For Cramer's V for 1 degrees of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect. For Cramer's V with 3_{df}; -.06=small effect, 0.17 = moderate effect, and 0.29 = large effect. All effect sizes significant (p<.001) unless stated otherwise.

Table 8: Two-step Solution 3: Two-step solution using age, healthiness, and number of vascular diseases and non-modifiable risk factors

	Overall	Cluster 1 n=35,290 (29.6%)	Cluster 2 n=32,700 (27.5%)	Cluster 3 n=33,013 (27.7%)	Cluster 4 n=18,077 (15.2%)	Effect Size
Continuous variables/counts	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**ϖ**
Variables used for clustering						
Age in years	48.5 (14.1)	33.9 (9.8)	50.3 (9.3)	56.3 (10.6)	59.7 (20.6)	.223
Median age	50	32	51	56	60	
Lifestyle healthiness score	28.9 (3.9)	26.2 (4.1)	29.9 (2.9)	31.0 (2.7)	28.7 (3.7)	.484
Number of vascular diseases	0.6 (1.0)	0.1 (0.3)	0.5 (0.5)	0.3 (0.4)	2.5 (0.9)	.698
Number non-modifiable risk factors	2.1 (1.5)	1.9 (1.1)	3.1 (1.0)	0.8 (0.7)	2.8 (1.4)	.410
Variables not used in clustering but related						
Number of modifiable risk factors	2.6 (1.4)	3.4 (1.3)	2.3 (1.2)	1.9 (1.1)	2.7 (1.3)	.190
Total number health concerns	5.3 (2.5)	5.5 (2.1)	5.9 (1.6)	2.9 (1.4)	7.9 (2.2)	.674
Categorical variables	**Overall**	**%**	**%**	**%**	**%**	**Cramer's V**
Risk factors, vascular diseases and age groups related to clustering variables						
Physical inactivity	43.7	60.7	39.3	26.5	49.9	.269
Smoking	12.6	21.3	9.1	7.1	11.7	.177
Excess alcohol	24.1	30.9	20.4	19.9	25.0	.111
Fatty food	13.1	21.4	10.4	7.9	11.2	.163
Fast foods	2.9	6.2	1.6	0.9	2.5	.131
Low fruit/vegetable	42.1	57.2	36.8	30.3	43.8	.217
Frequent stress	19.8	32.0	18.0	8.8	19.0	.222
Low fish consumption	53.2	65.9	51.2	42.8	51.2	.179
≥1 bad diet behaviour	69.6	84.7	66.0	57.3	69.1	.232
Overweight/obese	61.8	63.8	62.2	52.2	74.9	.150
Salt	27.5	48.9	19.7	16.5	19.9	.313
Unwilling to change:						
Inactive	52.8	64.1	44.4	41.3	51.6	.192
Smokers	37.8	47.1	27.4	23.0	36.0	.207
Excess alcohol	45.2	49.4	42.9	39.8	46.3	.077
Overweight	41.4	56.8	33.3	32.6	39.4	.213
Poor diet	24.9	39.5	16.1	13.1	23.0	.263
Stressed	32.3	45.2	18.7	14.9	27.8	.276
High salt	26.1	34.2	15.3	18.9	17.2	.197
Family history premature heart disease	37.4	31.0	60.7	13.6	51.5	.386
Family history of premature stroke	15.8	13.8	27.0	3.4	21.9	.251

	Overall	Cluster 1 n=35,290 (29.6%)	Cluster 2 n=32,700 (27.5%)	Cluster 3 n=33,013 (27.7%)	Cluster 4 n=18,077 (15.2%)	Effect Size
Family history dyslipidemia	44.6	42.8	67.1	13.2	64.9	.439
Family history hypertension	58.0	53.2	83.8	27.0	77.5	.460
Family history diabetes	44.0	44.8	66.4	17.5	56.0	.379
Higher risk ethnicity	5.5	6.9	7.8	1.5	5.8	.112
Diabetes	6.8	0.9	2.4	1.5	36.2	.493
Heart disease	4.4	0.3	1.0	1.2	24.3	.410
Hypertension	26.0	5.7	25.4	15.3	86.6	.607
Dyslipidemia	20.8	4.2	16.5	8.9	82.6	.655
Renal disease	1.3	0.2	0.4	0.2	6.8	.210
Stroke	2.1	0.2	0.7	0.6	11.0	.267
Variables not related to the clustering variables						
Mood disorder	16.9	19.9	17.2	10.9	21.8	.107
Prescribed medication	42.3	26.1	44.2	34.6	84.7	.388
Most/some of time miss medication†	12.4	20.0	12.3	9.1	10.2	.117
Proportion of those with diagnosis who have condition controlled "most of the time"						
Blood pressure	61.1	34.5	56.1	62.8	66.9	.111
Blood lipids	48.8	24.2	37.6	46.5	56.3	.130
Blood sugar	61.2	43.6	64.1	73.3	60.7	.067
Demographics						
Age groups:						
18-34	19.2	57.2	5.1	2.2	1.1	.443
35-44	18.1	26.8	21.6	11.5	6.7	
45-54	25.9	12.9	39.0	28.9	22.1	
55-64	24.1	2.9	28.2	35.4	37.6	
65-74	10.3	0.2	5.8	17.7	24.7	
75-90	2.4	0	0.2	4.2	7.9	
Age ≥65	12.7	0.2	6.0	21.9	32.6	.358
Entry portal/source:						
HRA page	11.6	10.8	11.8	12.4	11.5	.070
Mobile phone app	2.4	2.5	2.4	2.4	2.2	
BPAP	13.2	8.5	15.2	13.1	19.1	
eSupport	60.7	65.8	57.3	62.3	54.0	
HWAP	12.0	12.3	13.3	9.9	13.2	
Air Miles	60.3	65.3	56.9	62.0	53.5	.088
Enrollment:						
eSupport emails	30.7	34.4	29.8	29.0	28.4	.052
BPAP self-mgmt	1.8	0.8	2.1	1.4	3.7	.073
HWAP	7.3	7.1	8.4	5.9	8.4	.040
Joined any	38.4	41.4	38.7	35.1	38.3	.049
Male gender	32.0	29.6	25.0	34.0	46.0	.145
Higher education	76.6	80.1	78.0	76.7	67.0	.101

	Overall	Cluster 1 n=35,290 (29.6%)	Cluster 2 n=32,700 (27.5%)	Cluster 3 n=33,013 (27.7%)	Cluster 4 n=18,077 (15.2%)	Effect Size
Married	58.3	41.9	64.4	66.6	64.0	.216
Work full/part-time	58.5	68.5	65.3	51.4	39.9	.212
White collar occupation	65.0	66.7	69.3	64.7	54.7	.098

Missing=1,424 † Of those prescribed ≥1 medication * related to clustering variable(s)

For ϖ standard cut-offs are: 0.01 = small effect, 0.06 = medium effect, and 0.14 =large effect. For Cramer's V for 1 degrees of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect. For Cramer's V with 3_{df}; -.06=small effect, 0.17 = moderate effect, and 0.29 = large effect. All effect sizes significant (p<.001) unless stated otherwise.

Approach 4: Modifiable and non-modifiable risk factors as nominal variables

Table 9: Conditional probabilities of group membership by clustering variables, LCA Solution 1

	Cluster1	Cluster2	Cluster3	Cluster4
Cluster Size	0.3223	0.2619	0.2228	0.1929
Median age of group	45 yrs.	47 yrs.	52 yrs.	52 yrs.
Fruit & vegetables				
Healthy	0.8035	0.8174	0.2544	0.2520
Unhealthy	0.1965	0.1826	0.7456	0.7480
Fish consumption				
Healthy	0.6439	0.6379	0.2329	0.2132
Unhealthy	0.3561	0.3621	0.7671	0.7868
Salt consumption				
Healthy	0.8902	0.8587	0.4994	0.5281
Unhealthy	0.1098	0.1413	0.5006	0.4719
Family history dyslipidemia				
No family history	0.4295	0.7379	0.6491	0.3947
Family history	0.5705	0.2621	0.3509	0.6053
Family history premature heart disease				
No family history	0.3288	0.9883	0.9926	0.1976
Family history	0.6712	0.0117	0.0074	0.8024

Table 10: LCA Solution 1: Groups based on fruit and vegetable, fish and salt consumption and family history of dyslipidemia or premature heart disease

Variables	**Overall**	**Cluster 3 n=28,710 (24.1%)**	**Cluster 4 n=17,146 (14.4%)**	**Cluster 2 n=43,113 (36.1%)**	**Cluster 1 n=30,295 (25.4%)**	**Effect Size**
Continuous variables	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**ϖ**
Age (years)	48.6 (14.1)	44.6 (14.4)	46.2 (13.2)	50.5 (14.2)	51.0 (13.3)	.612
Median age	50	45	47	52	52	
Number vascular diseases*	0.6 (1.0)	0.5 (0.9)	0.8 (1.1)	0.5 (0.9)	0.8 (1.0)	.149
Number modifiable risk factors*	2.6 (1.4)	3.4 (1.2)	3.6 (1.2)	2.0 (1.2)	2.2 (1.2)	.503
Number non-modifiable risk factors*	2.1 (1.5)	1.4 (1.2)	3.3 (1.2)	3.1 (1.2)	1.3 (1.1)	.599
Health concerns*	5.3 (2.5)	5.3 (2.0)	7.5 (2.2)	3.8 (2.0)	6.1 (2.2)	.530
Overall healthiness score*	28.9 (3.9)	26.8 (3.8)	26.6 (3.8)	30.0 (3.3)	30.5 (3.2)	.456
Categorical variables	**Overall %**	**%**	**%**	**%**	**%**	**Cramer's V**
Variables used for clustering						
Fruit and vegetables	42.2	83.7	84.9	30.6	23.6	.680
Fish	53.2	88.4	89.7	32.7	28.4	.566
Salt	27.5	60.0	56.7	9.0	6.5	.554
Family history dyslipidemia	44.8	38.0	59.5	32.0	61.0	.261
Family history heart disease	47.6	0	100	0	91.4	.956
Variables not used for clustering						
Physical inactivity	43.7	54.1	58.2	34.0	39.5	.197
Smoking	12.6	17.0	18.1	9.1	10.2	.116
Alcohol	24.1	25.2	24.1	23.8	23.4	.016
Frequent stress	19.7	22.6	30.7	14.1	18.9	.140
≥1 bad dietary behaviour*	69.7	100	100	47.7	55.4	.525
Overweight/obese	61.8	62.0	67.5	58.3	63.5	.064
Unwilling to change:						
Inactive	52.8	57.2	55.3	49.7	48.9	.072
Smokers	37.9	42.5	42.5	32.8	32.5	.101
Excess alcohol	45.2	45.3	46.4	45.0	44.6	.011
Overweight	41.5	50.2	48.2	36.5	35.8	.132
Poor diet	24.9	33.7	32.3	14.6	14.8	.213
Stressed	32.3	40.7	37.8	25.7	24.5	.154
High salt	26.0	28.1	22.8	26.7	22.3	.057
Diabetes	6.8	5.3	8.8	5.7	8.8	.063
Heart disease	4.4	2.4	6.8	2.9	7.2	.104
Hypertension	26.1	20.5	30.5	23.4.	32.7	.111
Dyslipidemia	20.8	16.6	25.6	16.7	28.0	.127
Renal disease	1.3	1.1	1.6	1.1	1.5	.018
Stroke	2.1	1.7	2.6	1.8	2.5	.028
Mood disorder	17.0	17.8	24.3	13.1	17.5	.097
Family history	15.8	10.4	27.3	9.0	24.1	.213

Variables	Overall	Cluster 3 n=28,710 (24.1%)	Cluster 4 n=17,146 (14.4%)	Cluster 2 n=43,113 (36.1%)	Cluster 1 n=30,295 (25.4%)	Effect Size
stroke						
Family history diabetes	44.9	39.4	58.6	36.6	54.4	.182
Family history hypertension	58.1	49.9	71.7	48.9	71.2	.219
Higher risk ethnicity	5.5	6.2	7.1	4.4	5.3	.044
Prescribed medication	42.4	37.1	48.0	38.6	49.5	.110
Miss medication most/some of time†	12.4	15.8	16.8	9.7	10.5	.091
Proportion of those with diagnosis who have condition controlled "most of the time"						
Blood pressure	61.6	52.7	51.9	66.7	65.1	.154
Blood lipids	48.8	41.7	40.7	53.8	52.8	.082
Blood sugar	61.2	53.0	49.6	68.3	65.9	.103
Demographics						
Age Group:						
18-34 years	19.1	27.8	21.3	16.5	13.5	
35-44 years	18.1	21.5	22.7	15.5	15.9	
45-54 years	25.9	24.4	27.9	25.1	27.4	.115
55-64 years	24.1	17.8	20.1	26.7	28.7	
65-74 years	10.3	6.8	6.7	12.9	11.9	
75-90 years	2.4	1.6	1.2	3.2	2.6	
≥65 years	12.7	8.5	7.9	16.2	14.5	.107
Entry portal/source:						
HRA landing page	11.6	10.1	11.8	11.7	12.9	.047
Mobile phone app	2.4	2.6	2.5	2.4	2.2	
BPAP	13.3	11.9	15.1	12.4	14.8	
eSupport	60.7	63.0	55.1	63.5	57.6	
HWAP	12.0	12.4	15.5	10.1	12.4	
Air Miles	60.1	62.5	54.7	63.1	57.2	.068
Enrollment:						
eSupport emails	30.7	31.5	28.9	31.3	30.1	.020
BPAP self-mgmt	1.8	1.5	2.4	1.5	2.2	.029
HWAP	7.3	7.5	9.7	6.1	7.6	.045
Joined any	38.5	39.4	39.5	37.6	38.3	.017
Male gender	32.0	38.1	30.9	31.7	27.3	.083
Higher education	76.6	74.5	72.3	79.4	77.0	.060
Married	58.3	53.2	55.4	60.9	61.0	.069
Work full/part-time	58.5	62.6	62.8	55.7	56.3	.067
White collar occupation	65.0	61.9	61.5	67.2	67.0	.055

Missing=1,246 † Of those prescribed ≥1 medication * related to clustering variable(s)

For ϖ standard cut-offs are: 0.01 = small effect, 0.06 = medium effect, and 0.14 =large effect. For Cramer's V for 1 degrees of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect. For Cramer's V with 3_{df}; -.06=small effect, 0.17 = moderate effect, and 0.29 = large effect. All effect sizes significant (p<.001) unless stated otherwise.

Table 11: Two-step Solution 4: four-group solution based on fruit/vegetable, fish and salt consumption and family history of dyslipidemia and premature heart disease

	Overall	Cluster 2 n=32,798 (27.5%)	Cluster 4 n=30,607 (25.7%)	Cluster 1 n=34,678 (29.1%)	Cluster 3 n=21,181 (17.8%)	Effect Size
Continuous variables	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**ϖ**
Age (years)	48.6 (14.1)	43.4 (14.1)	49.3 (13.7)	51.0 (13.9)	51.5 (13.0)	.231
Median age	50	43	50	52	53	
Number vascular diseases	0.6 (1.0)	0.5 (0.9)	0.7 (1.0)	0.5 (0.9)	0.8 (1.0)	.129
Number modifiable risk factors *	2.6 (1.4)	3.7 (1.2)	2.8 (1.0)	1.8 (1.1)	1.9 (1.2)	.555
Number non-modifiable risk factors *	2.1 (1.5)	2.0 (1.5)	2.8 (1.0)	1.4 (1.1)	1.9 (1.2)	.401
Health concerns*	5.3 (2.5)	6.2 (2.4)	5.6 (2.3)	3.7 (2.0)	5.8 (2.2)	.414
Overall healthiness score*	28.9 (3.9)	26.3 (3.9)	28.3 (3.3)	30.9 (3.0)	30.6 (3.1)	.488
Categorical variables	**Overall %**	**%**	**%**	**%**	**%**	**Cramer's V**
Variables used for clustering						
Fruit and vegetables	42.2	60.0	100	0	0	.854
Fish	53.2	66.6	61.8	40.7	40.6	.239
Salt	27.5	100	0	0	0	1.000
Family history dyslipidemia	44.8	45.3	46.0	35.8	56.7	.141
Family history heart disease	37.6	35.7	39.0	0	100	.686
Variables not used for clustering						
Physical inactivity	43.7	51.2	53.0	32.9	36.4	.182
Smoking	12.6	16.8	15.2	8.4	9.1	.113
Alcohol	24.1	27.0	24.3	25.2	21.8	.047
High fat foods	13.1	23.9	12.6	6.9	6.9	.210
Fast foods	2.9	7.1	2.4	0.7	0.6	.162
Frequent stress	19.7	26.5	20.7	13.6	18.0	.124
≥1 bad dietary behaviour*	69.7	85.0	100	44.3	44.0	.535
Overweight/obese	61.8	60.1	66.3	58.7	63.3	.062
Unwilling to change:						
Inactive	52.8	55.1	55.7	49.3	47.2	.069
Smokers	37.9	42.4	38.5	32.5	31.4	.089
Excess alcohol	45.2	45.3	46.7	44.3	44.1	.020
Overweight	41.5	49.3	46.1	35.0	32.7	.140
Poor diet	24.9	37.4	23.7	12.8	11.5	.232
Stressed	32.3	40.9	32.7	23.9	22.2	.163
High salt	26.0	-	-	-	26.0	-
Diabetes	6.8	5.5	8.0	5.9	8.8	.054
Heart disease	4.4	3.3	5.4	2.9	7.4	.081
Hypertension	26.1	19.2	30.8	24.2	33.0.	.122
Dyslipidemia	20.8	17.6	24.1	17.5	26.5	.093

	Overall	Cluster 2 n=32,798 (27.5%)	Cluster 4 n=30,607 (25.7%)	Cluster 1 n=34,678 (29.1%)	Cluster 3 n=21,181 (17.8%)	Effect Size
Renal disease	1.3	1.2	1.4	1.1	1.5	.013
Stroke	2.1	1.9	2.3	1.8	2.7	.024
Mood disorder	17.0	20.1	18.3	12.8	17.0	.076
Family history stroke	15.8	16.0	16.7	9.2	25.0	.145
Family history diabetes	44.9	45.1	45.9	37.7	55.1	.117
Family history hypertension	58.1	55.4	60.2	50.9	71.0	.140
Higher risk ethnicity	5.5	6.9	5.8	4.1	4.9	.048
Prescribed medication	42.4	37.0	46.3	39.3	50.0	.101
Miss medication most/some of time†	12.4	18.3	12.4	9.2	9.6	.109
Proportion of those with diagnosis who have condition controlled "most of the time"						
Blood pressure	61.1	48.2	59.1	68.0	67.0	.105
Blood lipids	48.8	37.9	47.6	54.6	55.0	.090
Blood sugar	61.2	47.6	58.6	69.6	68.6	.114
Demographics						
Age Group:						
18-34 years	19.1	30.2	16.5	15.3	12.0	
35-44 years	18.1	22.5	18.6	15.1	15.3	
45-54 years	25.9	23.8	27.1	25.6	28.1	.133
55-64 years	24.1	16.6	24.6	27.5	29.6	
65-74 years	10.3	5.7	10.6	13.3	12.2	
75-90 years	2.4	1.2	2.6	3.2	2.7	
≥65 years	12.7	6.9	13.1	16.5	14.9	.113
Entry portal/source:						
HRA landing page	11.6	11.5	10.7	11.8	13.1	.036
Mobile phone app	2.4	2.6	2.2	2.5	2.2	
BPAP	13.3	12.7	14.0	12.1	15.0	
eSupport	60.7	59.5	61.3	63.4	57.0	
HWAP	12.0	13.8	11.3	10.2	12.7	
Air Miles	60.1	59.0	60.9	63.1	56.6	.047
Enrollment:						
eSupport emails	30.7	30.1	31.4	31.4	29.9	.014
BPAP self-mgmt	1.8	1.7	2.0	1.5	2.1	.019
HWAP	7.3	8.0	7.4	6.3	7.8	.027
Joined any:	38.5	38.7	39.1	37.8	38.2	.011 (p=.011)
Male gender	32.0	36.2	36.6	29.3	23.5	.108
Higher education	76.6	74.5	77.0	72.3	79.4	.060
Married	58.3	53.2	61.0	55.4	60.9	.069
Work full/part-time	58.5	62.6	56.3	62.8	55.7	.067
White collar occupation	65.0	61.9	67.0	61.5	67.2	.055

Missing=1,246 † Of those prescribed ≥1 medication * related to clustering variable(s)

For ϖ standard cut-offs are: 0.01 = small effect, 0.06 = medium effect, and 0.14 =large effect. For Cramer's V for 1 degrees of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect. For Cramer's V with 3_{df}; -.06=small effect, 0.17 = moderate effect, and 0.29 = large effect. All effect sizes significant ($p<.001$) unless stated otherwise.

Table 12: Probability of group membership for diabetes, hypertension, dyslipidemia age > 55 years and gender, LCA Solution 2

	Cluster1	Cluster2	Cluster3	Cluster4
Cluster Size	0.5805	0.3192	0.0841	0.0161
Median age	47 yrs.	58 yrs.	38 yrs.	50 yrs.
High fat foods				
Infrequent	0.9333	0.9367	0.2947	0.2194
Frequent	0.0667	0.0633	0.7053	0.7806
Fast foods				
Infrequent	0.9970	0.9966	0.7535	0.6865
Frequent	0.0030	0.0034	0.2465	0.3135
Salt intake				
Controlled	0.7408	0.8395	0.2490	0.3483
No controlled/high	0.2592	0.1605	0.7510	0.6517
Hypertension				
No diagnosis	0.9983	0.2510	0.9237	0.0920
Hypertension	0.0017	0.7490	0.0763	0.9080
Dyslipidemia				
No diagnosis	0.9036	0.5781	0.9011	0.4313
Dyslipidemia.	0.0964	0.4219	0.0989	0.5687

Table 13: LCA Solution 2: Latent cluster analysis for dietary behaviours and family history

	Overall	Cluster 3 n= 8,457 (7.1%)	Cluster 1 n=80,869 (67.5%)	Cluster 4 n=1,113 (0.9%)	Cluster 2 n=29,369 (24.5%)	Effect Size
Continuous variables/counts	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**ϖ**
Age in years	48.5 (14.1)	39.1 (13.2)	46.4 (13.7)	49.2 (12.2)	57.2 (11.5)	.373
Median age	50	38	47	50	58	
Number of vascular diseases*	0.6 (1.0)	0.3 (0.5)	0.2 (0.5)	2.2 (1.1)	1.8 (1.0)	.730
Number of modifiable risk factors*	2.6 (1.4)	3.9 (1.1)	2.4 (1.3)	4.2 (1.1)	2.6 (1.3)	.294
Number non-modifiable risk factors	2.1 (1.5)	2.1 (1.5)	1.9 (1.4)	3.0 (1.4)	2.3 (1.4)	.191
Total number health concerns*	5.3 (2.5)	6.3 (2.1)	4.5 (2.1)	9.4 (2.3)	6.9 (2.3)	.460
Lifestyle healthiness score*	28.9 (3.9)	24.9 (4.0)	29.3 (3.8)	24.9 (4.0)	29.0 (3.6)	.270
Categorical variables	**Overall**	**%**	**%**	**%**	**%**	**Cramer's V**
Variables used for clustering						
High fat foods	13.1	88.1	6.4	95.7	6.6	.664
Fast foods	2.9	31.7	0	54.7	0.4	.572
High salt	27.5	90.8	24.4	83.2	15.7	.420
Hypertension	26.1	9.3	0	100	100	.984
Dyslipidemia	20.8	10.2	13.5	72.8	42.0	.331
Variables not used for clustering						
Inactivity	43.7	57.3	40.5	68.5	47.6	.109
Smoking	12.6	19.8	12.4	19.4	10.6	.068
Alcohol	24.1	23.0	23.3	25.9	26.7	.035
Fruit/vegetable	42.4	71.5	38.9	77.0	41.5	.181
Low fish	53.3	72.5	51.8	74.4	50.9	.116
≥1 bad dietary*	69.7	100	66.9	100	67.7	.194
Overweight/obese	61.9	64.5	56.9	82.5	73.9	.154
Stress	10.9	34.4	17.9	42.1	19.9	.118
Unwilling to change:						
Inactive	52.8	57.6	53.5	58.1	49.2	.050
Smokers	37.9	44.4	37.5	43.2	34.8	.055
Excess alcohol	45.2	41.6	44.9	51.4	46.6	.027
Overweight	41.5	53.7	41.8	51.1	37.3	.085
Poor diet	24.9	43.1	23.3	40.9	20.7	.151
Stressed	32.3	43.3	32.4	39.7	25.9	.109
High salt	26.0	29.1	27.8	20.3	14.4	.112
Mood disorder	17.0	22.3	15.4	35.4	19.1	.076
Diabetes	6.8	3.1	3.3	23.9	17.1	.245
Heart disease	4.4	1.6	2.2	14.5	11.0	.192
Renal disease	1.3	0.7	0.4	5.6	3.6	.128
Stroke	2.1	0.9	1.0	7.6	5.2	.134

	Overall	Cluster 3 n= 8,457 (7.1%)	Cluster 1 n=80,869 (67.5%)	Cluster 4 n=1,113 (0.9%)	Cluster 2 n=29,369 (24.5%)	Effect Size
Ethnicity	5.4	8.2	5.0	8.1	5.7	.037
Family history stroke	15.8	16.8	14.3	24.2	19.4	.063
Family history dyslipidemia	44.8	46.1	43.0	69.5	48.3	.067
Family history diabetes	45.0	48.2	43.1	58.0	48.7	.057
Family history hypertension	58.2	55.0	49.0	85.8	83.3	.300
Family history heart disease	37.6	35.7	34.6	52.4	45.9	.104
Medication	42.4	33.0	29.1	74.9	80.4	.448
Most/some of time miss medication	12.4	22.8	13.5	25.7	9.5	.110
Proportion of those with diagnosis who have condition controlled "most of the time"						
Blood pressure	61.6	38.1	(empty)	43.5	62.3	.080
Blood lipids	48.8	27.0	42.8	33.6	57.1	.108
Blood sugar	61.2	42.3	64.8	34.9	61.8	.102
Demographics						
Age Groups*						
19-34	19.2	41.4	22.5	12.5	3.8	.215
35-44	18.1	24.7	20.5	20.7	9.3	
45-54	25.9	20.0	26.7	32.1	25.1	
55-64	24.1	10.3	21.4	24.9	35.6	
65-74	10.3	2.9	7.4	8.4	20.5	
75-90	2.4	0.6	1.4	1.3	5.7	
Age $\geq$ 65 yrs *	12.7	3.5	8.8	9.7	26.2	.235
Entry portal/source:						
HRA	11.7	12.1	11.5	11.2	11.9	.119
Mobile	2.4	3.6	2.4	4.3	2.1	
BPAP	13.3	12.3	9.2	24.7	24.3	
eSupport	60.6	54.8	65.4	43.8	49.8	
HWAP	12.0	17.2	11.5	16.0	11.9	
Air Miles	60.2	54.3	64.9	43.2	49.4	.143
Enrollment:		39.2				
eSupport	30.8		32.7	25.9	25.6	.067
BPAP self-mgmt.	1.8	2.4	0.8	5.5	4.4	.118
HWAP	7.3	10.0	6.9	9.8	7.5	.032
Joined any	38.4	39.5	39.3	39.2	35.5	.033
Male gender	32.0	38.6	28.5	54.2	39.1	.115
Higher education	76.6	75.4	79.2	67.3	70.1	.093
Married	58.3	46.1	57.4	54.3	64.4	.092
Work full/part-time	58.5	65.4	61.7	59.5	47.8	.126
White collar occupation	65.0	60.8	67.7	52.8	59.4	.082

Missing=582 † Of those prescribed $\geq$1 medication * related to clustering variable(s)

For ϖ standard cut-offs are: 0.01 = small effect, 0.06 = medium effect, and 0.14 =large effect. For Cramer's V for 1 degrees of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect. For Cramer's V with 3_{df}; -.06=small effect, 0.17 = moderate effect, and 0.29 = large effect. All effect sizes significant ($p<.001$) unless stated otherwise.

Table 14: Two-step Solution 5: Means and proportions for four-group two-step solution based on dietary risk factors and diagnosis of hypertension or dyslipidemia

	Overall	Cluster 2 n=16,866 (14.1%)	Cluster 3 n=24,377 (20.3%)	Group 1 n=48,080 (40.1%)	Cluster 4 n=30,505 (25.5%)	Effect Size
Continuous variables/counts	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**ϖ**
Age in years	48.5 (14.1)	43.4 (14.0)	44.6 (14.1)	46.7 (13.4)	57.5 (11.1)	.378
Median age	50	43	45	48	58	
Number of vascular diseases*	0.6 (1.0)	0.6 (0.9)	0.5 (0.9)	0.1 (0.2)	1.6 (0.9)	.659
Number of modifiable risk factors*	2.6 (1.4)	3.4 (1.2)	3.5 (1.2)	2.0 (1.2)	2.3 (1.2)	.479
Number non-modifiable risk factors	2.1 (1.5)	2.2 (1.5)	2.0 (1.5)	1.8 (1.4)	2.5 (1.4)	.178
Total number health concerns*	5.3 (2.5)	6.2 (2.4)	6.0 (2.4)	3.9 (2.0)	6.4 (2.2)	.466
Lifestyle healthiness score*	28.9 (3.9)	26.7 (4.0)	26.7 (3.9)	30.3 (3.3)	29.7 (3.3)	.131
Categorical variables	**Overall**	**%**	**%**	**%**	**%**	**Cramer's V**
Variables used for clustering						
High fat foods	13.1	93.0	0	0	0	.959
Fast foods	2.9	0	31.7	0	0	.424
High salt	27.5	51.1	100	0	0	.908
Hypertension	26.1	23.7	19.0	0	74.3	.674
Dyslipidemia	20.8	19.2	17.7	0	57.0	.556
Variables not used for clustering						
Physical inactivity	43.7	55.5	48.6	36.9	44.1	.132
Smoking	12.6	16.1	15.7	11.3	10.1	.074
Low fruit/veg	42.2	62.0	55.5	32.0	36.6	.240
Low fish	53.3	67.7	64.2	47.1	46.3	.180
≥1 bad dietary *	69.7	100	79.6	59.5	61.2	.315
Frequent stress	19.8	31.6	23.4	15.8	16.1	.140
Excess alcohol	24.1	20.3	28.3	22.5	25.5	.063
Overweight/obese	61.9	68.0	58.1	56.4	71.7	.145
Unwilling to change:	52.8	56.7	54.1	53.1	48.7	.055
Inactive	37.9	43.7	41.6	34.3	34.5	.084
Smokers	45.2	41.6	46.0	44.9	46.6	.030
Excess alcohol	41.5	49.9	47.6	39.2	35.8	.110
Overweight	24.9	35.3	34.6	17.3	17.1	.203
Poor diet	32.3	39.1	39.1	28.0	23.5	.144
Stressed High salt	26.0	28.1	28.1	-	-	.029
Mood disorder	17.0	22.8	18.8	13.6	17.5	.085
Diabetes	6.8	6.1	5.6	2.5	15.2	.201
Heart disease	4.4	3.8	3.4	1.5	10.2	.169
Renal disease	1.3	1.4	1.2	0.4	2.7	.082
Stroke	2.1	2.0	1.9	0.8	4.3	.099

	Overall	Cluster 2 n=16,866 (14.1%)	Cluster 3 n=24,377 (20.3%)	Group 1 n=48,080 (40.1%)	Cluster 4 n=30,505 (25.5%)	Effect Size
High risk ethnicity	5.4	6.6	6.4	4.6	5.3	.036
Family history premature stroke	15.8	17.3	15.4	13.9	18.4	.051
Family history dyslipidemia	44.8	47.9	44.2	37.6	55.0	.140
Family history diabetes	45.0	48.1	43.9	42.7	47.7	.047
Family history hypertension	58.2	60.1	54.5	49.3	74.0	.201
Family history premature heart disease	37.6	38.3	35.0	33.6	45.7	.102
Medication	42.4	41.0	33.8	23.5	77.3	.434
Most/some of time miss medication †	12.4	18.4	16.5	12.2	9.0	.110
Proportion of those with diagnosis who have condition controlled "most of the time"						
Blood pressure	61.1	49.1	51.1	(empty)	65.2	.118
Blood lipids	48.8	36.7	40.9	(empty)	53.1	.109
Blood sugar	61.2	45.2	50.6	71.0	65.3	.116
Demographics						
Age group						
18-34	19.2	30.5	27.2	21.2	3.3	.219
35-44	18.1	22.2	22.0	20.7	8.6	
45-54	25.9	23.8	24.7	27.9	24.9	
55-64	24.1	16.6	18.2	21.4	37.3	
65-74	10.3	5.7	6.5	7.4	20.5	
75-90	2.4	1.2	1.4	1.4	5.5	
Age ≥65 yrs	12.7	6.8	7.9	8.8	26.0	.235
Entry portal:						
HRA	11.7	12.3	11.2	11.4	12.2	.087
Mobile	2.4	3.4	2.3	2.2	2.3	
BPAP	13.3	13.8	12.4	9.3	20.0	
eSupport	60.6	54.6	61.7	66.6	53.6	
HWAP	12.0	15.9	12.5	10.5	11.9	
Air Miles	60.2	54.0	61.2	66.2	53.2	.117
Enrollment:						
eSupport	30.6	29.3	30.7	33.1	27.6	.049
BP self-mgmt.	1.8	1.9	1.6	0.8	3.5	.082
HWAP	7.3	9.8	7.2	6.2	7.7	.046
Joined any	38.4	39.6	38.4	39.0	36.8	.021
Male gender	32.0	37.4	34.8	25.3	37.4	.120
Higher education	76.6	75.5	77.0	80.0	71.4	.081
Married	58.3	51.4	54.7	58.3	64.8	.091
Work full/part-time	58.5	62.7	62.5	62.4	46.9	.138
White collar occupation	65.0	62.0	64.4	69.2	60.6	.077

Msg = 682 † Of those prescribed ≥1 medication * related to clustering variable(s)

For ϖ standard cut-offs are: 0.01 = small effect, 0.06 = medium effect, and 0.14 =large effect.

For Cramer's V for 1 degrees of freedom (two categories), 0.01 = small effect, 0.30 = medium effect, and 0.50 = large effect. For Cramer's V with 3_{df}; -.06=small effect, 0.17 = moderate effect, and 0.29 = large effect. All effect sizes significant ($p<.001$) unless stated otherwise.